Crowdsourcing veridicality annotations in Spanish: How much do speakers disagree?

SAARLAND UNIVERSITY
Department of Language Science and Technology

Master Thesis

Author
Teresa Rosa Martín Soeder
Matriculation
2581298

Supervisors
Prof. Vera Demberg
Dr. Lucia Donatelli

June 21, 2023

UNIVERSITÄT
DES
SAARLANDES

Declaration of Authorship

Eidesstattliche Erklärung

Ich erkläre hiermit an Eides Statt, dass ich die vorliegende Arbeit selbstständig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel verwendet habe.

Statement in Lieu of an Oath

I hereby confirm that I have written this thesis on my own and that I have not used any other media or materials than the ones referred to in this thesis.

Einverständniserklärung

Ich bin damit einverstanden, dass meine (bestandene) Arbeit in beiden Versionen in die Bibliothek der Informatik aufgenommen und damit veröffentlicht wird.

Declaration of Consent

I agree to make both versions of my thesis (with a passing grade) accessible to the public by having them added to the library of the Computer Science Department.

Datum/Date: _____

Unterschrift/Signature: _____

SAARLAND UNIVERSITY
Department of Language Science and Technology

Abstract

Crowdsourcing veridicality annotations in Spanish: How much do speakers disagree?

by Teresa Rosa Martín Soeder

In veridicality studies, an area of research of Natural Language Inference (NLI), the factuality of different contexts is evaluated. This task, known to be a difficult one since often it is not clear what the interpretation should be Uma et al. [1], is key for building any Natural Language Understanding (NLU) system that aims at making the right inferences. Here the results of two studies that analyze the veridicality of mood alternation and specificity in Spanish, and whose labels are based on those of Saurí and Pustejovsky [2] are presented. The first one, the pilot study, has an inter-annotator agreement score of $AC_2 = 0.484$, slightly lower than that of de Marneffe et al. [3] ($\kappa = 0.53$), a main reference to this work; shows some mood-related significant effects, and presents a few unexpected tendencies, like high factuality despite the presence of a negation marker. The second one, the main study, has an inter-annotator agreement of $AC_2 = 0.114$, quite lower than the first study, and a couple of mood-related significant effects. Due to this strong lack of agreement, an analysis of what factors cause disagreement is presented together with a discussion based on the work of de Marneffe et al. [3] and Pavlick and Kwiatkowski [4] about the quality of the annotations gathered and whether other types of analysis like entropy distribution could better represent this corpus. The annotations collected for both studies are available at `https://github.com/narhim/veridicality_spanish`.

Acknowledgements

The list of who to thank is long, can words be enough? The obvious start is with both my supervisors, Lucia and Vera, thanks to their guidance *this* has come through. Then there is of course Toloka since without their grant I would not have been able to finance my study. Last, but not least, all the people whose support and advice have brought me here:

My father, whose path into academia I try to follow;

My mother, whose love for languages I inherit;

Each of my seven siblings, by blood or by law: three countries, two continents, one heart;

My friends back south, who never deterred me from trying new things;

And the friends from different corners of the world whom I have been lucky enough to encounter here.

The path has been long and not always easy, you have not left my side and I will not forget it. May this be one of the many journeys we will share together, may each of your dreams become real in the most splendid way, and may God hold you in the palm of His hand.

THANK YOU! ¡MUCHÍSIMAS GRACIAS! VIELEN DANK!

Content

Declaration of Authorship i

Abstract ii

Acknowledgements iii

1 Introduction 1
1.1 Introduction . . . 1

2 Background 7
2.1 Veridicality and Factuality . . . 7
2.2 Mood in Spanish . . . 12
2.2.1 Mood Alternation . . . 14
2.3 Specificity . . . 17

3 Pilot Study 21
3.1 Design . . . 21
3.2 Dataset . . . 24
3.3 Predictions . . . 24
3.4 Procedure . . . 26
3.5 Results . . . 27
3.5.1 Overview of the Annotations . . . 27
3.5.2 Inter-Annotator Agreement Scores . . . 30
3.5.3 Label Space Reductions . . . 32
3.5.4 Model Fitting . . . 33
3.5.5 Comparison with Predictions . . . 35
3.6 Discussion . . . 38

4 Main Study 41
4.1 Introduction . . . 41
4.2 Dataset . . . 43

4.3 Predictions . . . 44
4.4 Procedure . . . 45
4.5 Analysis of NaS . . . 47
4.6 Results . . . 48
4.6.1 Overview of the Annotations . . . 48
4.6.2 Inter-Annotator Agreement Scores . . . 49
4.6.3 Model Fitting . . . 51
4.6.4 Agreement Patterns . . . 54
4.6.5 Manual Analysis . . . 58

5 Conclusions **64**
5.1 Summary of Results . . . 64
5.2 Disagreement . . . 66
5.3 Answer to Research Questions . . . 69
5.4 Future Work . . . 70

List of Figures **72**

List of Tables **73**

Bibliography **74**

Additional Scores **79**
A.1 Pilot Study . . . 79
A.1.1 Inter-Annotator Agreement Scores . . . 79
A.1.2 Cumulative Link Mixed Model . . . 80
A.2 Main Study: CLMMs with Most Frequent Matrices . . . 82
A.2.1 Predictors: Mood . . . 82
A.2.2 Predictors: Mood and Matrix . . . 83

To those struggling to find their place in the world, keep fighting, it's worth the effort.
To the researchers forgotten by the institutions, your voice is heard and you are not alone.

Chapter 1

Introduction

1.1 Introduction

Is that true? Did it really happen? Often we ask ourselves or our interlocutors these questions, and we try to assess if the information conveyed is likely to be truthful or not, that is, if it corresponds to actual situations in the real world [2]. Also, as speakers or authors we usually try to portray what we know about the **truthfulness** or **factuality** of the events conveyed, which are here loosely defined as *anything that happens or is* like "being 3 years old" or "having bought a house". Therefore this kind of language understanding needs to be incorporated into any system that aims at comprehending human language.

Let us consider examples (1a) to (1c), extracted from the dataset SQuAD [5], where an interaction for a question-answering system is portrayed. To give such a simple response as "seven" the system is making an inference from the context given. Specifically, it is inferring that "Seven Monument Zones are present in the Kathmandu Valley" follows or *loosely* entails from "These are [...] in the seven well-defined Monument Zones of the Kathmandu valley". The term **follows** and **loosely entails** are used to denote that we are referring to pragmatic inferences and not logical entailment [6].

(1) a. Question: "How many Monument Zones are present in the Kathmandu valley?"

b. Context: These are amply reflected in the many temples, shrines, stupas, monasteries, and palaces in the seven well-defined Monument Zones of the Kathmandu valley are part of a UNESCO World Heritage Site.

c. Answer: "seven"

But, how does the question-answering system or we as speakers understand that the inference is correct, or in other words, that it is true that there are seven monument zones in the Kathmandu Valley? There are different linguistic factors at play, like the use of negation or modality markers exemplified both in Spanish and in English in utterances (2a) to (2c). All three examples have one event in common "Pedro has done the laundry" (*Pedro ha hecho la colada*), but in example (2b) we have the negation adverb "not" (*no*) and in example (2c) we have the modal verb "must" (*tener que*). The question now is whether the event "Pedro has done the laundry" can be inferred from each of these utterances.

In the first case, the answer is a clear yes, unless there is something in our **world knowledge** about Pedro that lead us to not be certain about it or to completely negate its veracity. In example (2b) the situation is the exact opposite. Due to the presence of the negation adverb "not" (*no*) the most likely situation is that we negate the truthfulness of "Pedro has done the laundry", unless, yet again, we know something about Pedro that changes this assessment. As to utterance (2c), the situation is roughly in the middle between the two previous sentences. The presence of the deontic modal "must" (*tener que*) modifies the truthfulness or factuality of the event. It does not make it completely true, but it does not make it completely false. It instead yields the "Pedro has done the laundry" (*Pedro ha hecho la colada*) an obligation [7], puts it in the realm of possibilities, and thus makes its factuality uncertain and more dependent upon the listener than the two previous utterances, both for Spanish and English speakers.

(2) a. Pedro ha hecho la colada.
Pedro have.PRS.IND.3SG do.PTCP the.F.SG laundry.M.SG
"Pedro has done the laundry."

b. Pedro no ha hecho la colada.
Pedro not have.PRS.IND.3SG do.PTCP the.F.SG laundry.M.SG
"Pedro hasn't done the laundry."

c. Pedro tiene que haber hecho la colada.
Pedro tiene.PRS.IND.3SG that have.INF do.PTCP the.F.SG laundry.M.SG
"Pedro must have done done the laundry."

The study of what linguistic factors affect the factuality of an event, like the analysis of "not" and "must" that has just been presented, is called veridicality, the research topic for this thesis. To be more specific, veridicality is an area of research within natural language inference (NLI) and theoretical linguistics that studies the truth value of a proposition or event in a specific context [8, 9]. As to NLI, it is a branch of natural

language understanding (NLU) with its main task being entailment classification, that is, as it has been done above, given premises like utterances (2a) to (2c) in Spanish or English, and the hypothesis "Pedro has done the laundry" (*Pedro ha hecho la colada*), the task is to classify the relationship between each premise with the hypothesis by picking a label from a usually small set of labels like {*entailment*[1]*, neutral, contradiction*} [10] or {*yes, unknown, not*}, depending on how the task is defined. In our case, the most likely scenario is that we pick the labels yes, not, and unknown respectively. For a system to successfully complete this task it needs to form a thorough and complete meaning representation of both sentences [10], and here is where veridicality, among other disciplines, comes into play.

The focus of this thesis is the analysis of veridicality judgments in Spanish; that is, the analysis of factuality judgments in specific contexts in Spanish. In particular, the goal is to analyze how mood alternation, in other words, the possibility of using a verb either in indicative or subjunctive mood; and the specificity of the syntactic subject, that is, the identifiability of the referent in the discourse universe [11], affect factuality judgments about an event. These factors are presented in utterances (3a) and (3b).

Assuming for the moment that the alternative subjects in examples (3a) and (3b) are one entity, we see that these two sentences are almost identical. The only difference lies in the embedded verb *tener* (to have). As indicated by the glosses, in the first case the verb is in the indicative mood, and in the second one the verb is in the subjunctive mood. Whereas normally a verb is only allowed to be either in the indicative or in the subjunctive mood, here both possibilities are allowed and this is what is called mood alternation. More details about this phenomenon will be given in Chapter 2, but it should be noted now that there is no straight translation of this difference into English, thus translations are kept exactly the same, as in Faulkner [12].

(3) a. El gobierno/ El presidente no dijo
the.M.SG government/ the.M.SG president.M.SG not say.PST.PFV.IND.3SG
que el país tenía problemas
that the.M.SG country.F.SG have.PST.IPFV.IND.3SG problem.M.PL
económicos.
economic.M.PL
The government/ The president didn't say that the country had economic problems.

[1] *In a loose sense.*

b. El gobierno/ El presidente no dijo
the.M.SG government/ the.M.SG president.M.SG not say.PST.PFV.IND.3SG
que el país tuviera problemas
that the.M.SG country.F.SG have.PST.IPFV.SBJV.3SG problem.M.PL
económicos.
economic.M.PL
The government/ The president didn't say that the country had economic problems.

As to the alternative subjects *el gobierno / el presidente* (the government / the president), they reflect differences in specificity. The first option, *el gobierno* (the government), refers to a set of people who rule or administer a country, that is, it refers to a set of entities. In opposition to this, there is the second option, *el presidente* (the president), which refers to a single entity. As with mood alternation, more details about this phenomenon and how it affects the factuality of an event are provided in the next chapter.

The above-mentioned goal of understanding how mood alternation and specificity affect the factuality of an event is realized in the following research questions:

RQ1.- In a complex sentence, how does the mood alternation of the embedded verb that occurs due to the negation of the main or matrix verb affect the factuality value of the embedded event?

RQ2.- In a simple sentence, how does the mood alternation caused by an adverb of doubt or possibility affect the factuality judgment of the event?

RQ3.- How does an individual subject affect the factuality judgment of the event?

RQ4.- How does a subject that refers to a collective entity like an institution, affect the factuality judgment of the event?

To answer these questions and verify their relevance, to also verify experimental soundness, and be able to get direct feedback from the annotators, a pilot study was run among friends and relatives who are linguistically-naive native Spanish speakers. Given the results obtained which are introduced in Chapter 3, it was concluded that answering these four questions required a complicated experimental design and analysis. Thus it was decided to remove research question RQ2.- and for the main study only consider the following three:

RQ1.- In a complex sentence, how does the mood alternation of the embedded verb that occurs due to the negation of the main or matrix verb affect the factuality value of the embedded event?

RQ2.- How does an individual subject affect the factuality judgment of the event?

RQ3.- How does a subject that refers to a collective entity like an institution, affect the factuality judgment of the event?

For the main study, a similar methodology to the pilot experiment was followed: annotations were gathered from linguistically-naive native Spanish speakers. But instead of acquaintances, annotators were recruited from the Toloka platform [13]. Consequently, the sociolinguistic background increased significantly from mostly speakers from Spain to speakers from all countries where Spanish is either an official language or an important minority language.

Regarding the motivations behind this thesis, firstly Spanish was chosen not just for being my mother tongue, but because, to my knowledge, there is a considerable lack of veridicality studies and available corpora. As will be explained in the next chapter, most of the studies found were done from a different perspective since annotations informed not about the relation between a premise and a hypothesis, but directly annotated lexical elements with factuality labels. Even more, most of these corpora are annotated by linguistic experts. As it was mentioned above, here instead, the aim was to gather annotations from linguistically naive speakers to obtain pragmatically informed labels as is often the case in natural language processing (NLP). More details about the differences between these perspectives will be given in the next chapter.

The most practical contribution from this thesis is two annotated corpora available at `https://github.com/narhim/veridicality_spanish`. For both of them the raw annotations are given, that is, instead of the aggregated label for each pair, the label assigned by each annotator is offered. In addition, every supplementary information, like translations or morphological annotations, that has been generated is also given. The first dataset, resulting from the pilot study, consists of 306 pairs annotated by 5 annotators each. The second one, resulting from the main study, consists of 477 pairs annotated by 7 annotators each.

Another important contribution is the analysis presented. It is done from a statistical perspective and a linguistic one, with the idea of not *just* answering the research questions, but trying to understand the data as much as possible within the limitations of

this thesis. This gives us a more complete picture of the annotations gathered and might help future researchers when trying to decide how to understand their data.

Next, Chapter 2 introduces the literature review and the explanation of the main concepts used here. Then, Chapter 3 presents the already mentioned pilot study and then the main study is introduced in Chapter 4. Finally, Chapter 5 brings together all the results and issues discussed in each experiment, suggests possible improvements, and proposes some lines of future work.

Chapter 2

Background

The goal of this chapter is to clarify the most important terms that will be used in this thesis, give theoretical support for the claims made for the experiments, and present some previous research important to the study here. Especially the focus is on distinguishing concepts that are often confused, like veridicality and factuality, and introducing the reader to the phenomena in Spanish linguistics on which this thesis focuses.

First Section 2.1 clarifies the difference between veridicality and factuality, provides specific work on each of them, and explains the two main approaches used to study these concepts. Secondly, in Section 2.2, the first experimental condition in this thesis will be explained, mood in Spanish, together with the specific mood contexts set as experimental conditions. After this, in Section 2.3 the second experimental condition, specificity, is introduced, together with the particular specificity contexts whose veridicality is here analyzed.

2.1 Veridicality and Factuality

Let us consider examples (4a) to (4d), where we have events that are intrinsically related. If we were to do an NLI study with these examples, we could directly study the *thruthfulness* of each of them as a single event, i.e., we could study the **factual nature** of each example towards the real world or the events in the discourse [2]. Another option would be to study the factual nature of the event *Anna's father has arrived* in the different contexts in which is presented: standing completely on its own (4a), or as part of a complex event (4b) to (4d). In this case, the goal would be not to understand the factuality of *Anna's father has arrived*, but rather to understand how its factuality changes when the event is embedded under an epistemic verb (4b), a verb of believe

(4c), and a verb of speech (4d). The former case is a factuality study and the latter a **veridicality study**, as in the experiments presented here.

(4) a. Anna's father has arrived.
b. John knows that Anna's father has arrived.
c. John believes that Anna's father has arrived.
d. John says that Anna's father has arrived.

Probably one of the most important studies in NLI with a factuality focus is that of Williams et al. [10]. They extended the work of Bowman et al. [14], which presented an NLI corpus where the premises were crowdsourced figure captions and hypotheses were also crowdsourced, by increasing the number of genres to a total of 10: transcribed conversations, official documents, letters, the public report of 9/11, non-fiction works, popular cultural articles, telephone transcriptions, travel guides, short posts about linguistics, and fiction works. This corpus, denominated as MNLI or Multi-NLI, uses the labels {*entailment, unknown, contradiction*} and is now an important benchmark in NLI, as proven, for example, by its use to test BERT [15]. Here, as in this corpus, annotations for premise-hypothesis pairs are crowdsourced. Furthermore, from this corpus originates the XNLI corpus [16], which consists of translations of the MLNI corpus into different languages, including Spanish. As we will see in Chapter 4, a small subset of this corpus was used in the main study.

Another important work in NLI is the FactBank corpus [2], which consists of $9,488$ manually annotated events by experts which resulted in the FactBank corpus. This means that instead of annotating premise-hypothesis pairs as in the standard NLI task, they annotate events within the sentences they occur. It can be said that this work is done from a factuality focus since it is not designed following different contexts, but the authors indeed present an extensive analysis of the different factors that affect the factuality of an event, that is, they present a veridicality analysis. Another important feature of their work is their set of labels, which is the baseline of the one used here, and which results from the combination of an epistemic scale, {*certain, likely (probable), possible*}, with a quality scale, {*positive, negative*}. They map this combination to the traditional Square of Opposition, but for the sake of simplicity, here this square is reduced to a linear scale of factuality, as we will see later on. In any case, the labels they used are the following: *certainly yes* (CT+), *probably yes* (PR+), *possibly yes* (PS+), *certainly not* (CT-), *probably not* (PR-), *possibly not* (PS-), *unknown or uncommitted* (Uu), and

certain but unknown output (CTu).

An interesting factuality study is that of Pavlick and Kwiatkowski [4]. Their goal was to determine whether the disagreement often seen in NLI datasets is noise or a reproducible signal, and thus it should be included in data analysis and modeling. To fulfill this goal, they collected factuality judgments on 500 pairs from different corpora with 50 annotators per pair. But after filtering the annotations, 496 pairs with a mean of 39 workers per pair were left to analyze.

When analyzing the annotations, the authors assumed that if the disagreement is noise, the gathered labels can be modeled as a simple Gaussian distribution where the mean is the true label and, consequently, there is only one true label. To verify this assumption, the authors fixed two models for each pair: one single Gaussian and a Gaussian Mixture Model (GMM) where the number of components is chosen during training. Results showed that overall there was a better fit with GMMs and that for 20% of the pairs, there was a nontrivial second component, that is, that for 20% of the pairs, there is not just one *true* label, but rather two. This phenomenon which can be referred to as **label split** [3] and be caused by **inherent disagreement**, might help explain the results obtained in the main study.

Furthermore, Pavlick and Kwiatkowski [4] analyzed whether context reduces disagreement by collecting annotations at three levels: word, sentence, and paragraph. Results showed that disagreement increases with context. They explain this by hypothesizing that less context may result in higher agreement because with less context humans can more easily call upon *default* interpretations. This supports the decision made here of not including out-of-sentence context in either of the experiments presented here.

A clear example of a study on veridicality is that of Ross and Pavlick [17]. Their goal was to learn whether neural models with no explicit knowledge of verbs' lexical categories can make inferences about veridicality consistent with human inferences. To do so, they crowdsourced human judgments on 1,500 sentences with 137 verb complement constructions using as labels a 5-point Likert scale, and compared these judgments with those made by BERT. Their results and analysis are quite thorough, and led them to conclude that although BERT was able to replicate many of the human judgments, there is still significant room for improvement. Important from this work is also their explanation on the different perspectives that veridicality and factuality studies can have: lexical semantics or sentence meaning approach, and pragmatic or speaker meaning approach.

But before explaining their definition of these approaches, we should go over the work of de Marneffe et al. [3].

The study of de Marneffe et al. [3] is a main reference to this thesis. Their goal was to identify some of the linguistic and contextual factors that shape reader's veridicality judgments. To do so, they present crowdsourced annotations on a part of the FactBank corpus with the same set of labels minus one, CTu, and a system for veridicality assesment. For our purposes, the two most important parts of their work is the consideration of the possible occurrence of label split amd the comparison they made between their annotations and the original FactBank's annotations.

In order to see if there is a possibility of label split, they studied the plotted distribution of **agreement patterns**, that is, of how votes for the different labels are distributed in each pair. After discarding the likely noisy patterns, they observed that they are many examples for which is unlikely that the agreement pattern is due to noise. For example, for the premise "In a statement, the White House said it would do "whatever is necessary" to ensure compliance with the sanctions." which had as a hypothesis "there will be compliance with the sanctions", votes were equally split between the labels Uu and PR+, that is, it had an agreement pattern of $[5 - 5]$. The authors explain this difference by stating that the judgment in this case depends heavily on the speaker's previous knowledge about the White House.

The comparison they make between the two set of annotations is very important because it shows quantitative differences between what they called, a lexical approach and a pragmatic approach. They do not define in detail the **lexical approach**, but they do mention that in such approach the context in which we study the factuality of a proposition is a lexical item. Ross and Pavlick [17] extends this definition by adding that a system following it should aim to model the aspects of a sentence semantics, thus, this representation can be derived from the lexicon and is independent of context, and also, as stated in Saurí and Pustejovsky [2], such a system would no consider world knowledge. Thus, linguistic experts are usually the ones who annotate the data.

As to the **pragmatic approach**, approach used by de Marneffe et al. [3], by Ross and Pavlick [17], and as hinted in the previous chapter, it is also used here. This approach requires the system to derive a representation of the sentence that considers the communication intent for that sentence in a specific context, that is, a goal-directed representation of a sentence within the context it was created [17]. Such a representation

entails two important things: the consideration of world knowledge, and the embracing of uncertainty [3]. To obtain this representation a key step is to collect annotations from linguistically naive workers, so for examples (4a) to (4d) they would consider any knowledge they might have about John, Anna, and her father; and they would ignore any notions as to what *to say* is *supposed* to mean, and instead just use their linguistic intuition.

There are other two important facts to consider about these two approaches. The first one is that not everyone explicitly defines the approach they use, but it is often, if not always, possible to infer it. Considering this is very important, specially when comparing results from different studies. The second fact is how these approaches are referred to. de Marneffe et al. [3] and Ross and Pavlick [17] use the terms lexical and pragmatic, but Ross and Pavlick [17] also employs the terms sentence meaning approach and speaker meaning approach respectively. Furthermore, de Marneffe et al. [3] hints at the idea that the pragmatic approach is done from the reader's persective, since we want to analyze what the reader understands, not what the author says, as in the lexical approach [2]. Some studies and tasks like FACT at the Iberian Languages Evaluation Forum (IberLEF) [18], make use of these terms, author's or reader's perspective, rather than semantic or pragmatic approach.

Aside from the already mentioned Spanish section of the XNLI corpus, there are two other corpora that should be mentioned here: SenSem [19] and TAGFACT [20]. Both follow Saurí and Pustejovsky [2] in annotating events within the sentence they occur instead of annotating premise-hypothesis pairs and having experts instead of lay-workers as here doing the annotation. The difference between the two corpora is that TAGFACT focuses on factuality annotations, whereas SenSem codifies information regarding aspectuality, modality, polarity and factuality.

Now that we have seen the differences between a factuality and veridicality focus, and between a lexical and a pragmatic approach, we can go on to exploring another important topic in this thesis: mood in Spanish.

2.2 Mood in Spanish

Simply put, **mood** is the grammaticalization of modality [21, 22], and thus it has been traditionally related with the speaker's attitude towards an utterance [21, 23]. Although this notion is imprecise [23] and more needs to be said in order to explain all the phenomena, for our purposes is enough.

Since the commitment of the speaker usually takes form in different degrees [21], in most languages mood takes form in different subcategories, like the indicative, the subjunctive and the imperative in Italian, or the subjunctive and the conditional in Hungarian. For Spanish, as in Italian, nowadays most of the grammarians agree on the existence of three subcategories of mood[1]: indicative, subjunctive and imperative. Since the imperative mood is out of the scope of this thesis, I would simply say that it is the mood mainly used to express commands. As to the indicative and subjunctive mood, the topic is not so clear.

Quite often, the indicative and subjunctive moods are defined in opposition to each other Lyons [21], and Spanish is not an exception. Some pairs of concepts that Real Academia Española [23] uses to describe this opposition are the following: certainty/uncertainty, reality/virtuality, actuality/non actuality and commitment of the speaker with the veracity of what is spoken/lack of commitment. So for example, in (5a), where the verb *estar* (to be) is in indicative, the reading that we get is that *it is a reality that Sofía was there to see it.* Contrary to this, in (5b), where the same verb, *estar*, is in its past perfect tense from the subjunctive, the reading is that *Sofía wasn't there, but we wish there was world, a possible world, in which she was there.* But, as stated in Real Academia Española [23] and Sánchez-Jiménez [22], these oppositions do not always work well.

(5) a. Sofía **estuvo** allí para verlo.
Sofía **be.pst.pfv.ind.3sg** there to see.INF.it
"Sofía was there to see it."

b. ¡Si Sofía **hubiera estado** allí para verlo!
if Sofía **have.pst.pfv.sbjv.3sg be.ptcp** there to see.INF.it
"If only Sofía had been there to see it!"

[1] For a diachronic review of the study of mood in Spanish see Calvo [24]

(6) a. Quiero suponer que **has preparado**
want.PRS.IND.1SG assume.INF that **have.prs.ind.2sg prepare.ptcp**
todo.
everything
"I want to assume that you **have prepared** everything."

b. Siento mucho que se te
feel.PRS.IND.1SG a.lot that itself you.DAT
haya averiado el coche.
have.prs.sbj.3sg break.down.ptcp the.M.SG car
"I'm so sorry to hear that your car **broke down**."

For example, in (6a), where the embedded verb *preparar* (to prepare) is in indicative, the reading for the event is *I want the fact that you have prepared everything to be a reality, but it might not be*; and, opposite to this, in (6b), where the embedded verb *averiarse* (to break down) is in the subjunctive mood, its reading is that *it is a reality that your car broke down and I feel sorry for that.* This does not mean that the above mention oppositions are useless, but it does imply that interpreting them too strictly would be a mistake, and that some complimentary considerations are necessary. In this regard, Villalta [25] talks about an ordering relation between contextual alternative propositions as the cause for embedded propositions to be in the subjunctive mood, and Mejías-Bikandi [26] explains the contrast in terms of old and new information.

Specifically, Mejías-Bikandi [26] understood **old information** as the information that is pragmatically presupposed, or in other words, the information with which the speaker assumes familiarity. **New information** would be then the one that is not presupposed. Based on these definitions, he classifies **matrix verbs**, that is, verbs that take a verbal predicate as a complement, into two groups: those that introduce old information and those that introduce new information. Then he uses this classification to explain the distribution in mood in complements of different types of matrices. The indicative, he claims, is used when the matrix introduces new information, as in the case of mental matrices like *notar* (to notice); and the subjunctive is instead used when the matrice introduces old information, as in the case of comment matrices like *lamentar* (to regret). Furthermore, he uses this classification to explain the distribution of mood for the following phenomena: the negation of some matrix verbs and matrix verbs whose subject contain a quantifying expression such as *poco/a/s* (a bit/bits) or the adverb *solo* (only). Next, we will briefly go over the syntactic functioning of the indicative and subjunctive moods.

As stated in Real Academia Española [23], authors often talk about a **dependent** and an **independent mood**, the first one being the one that requires a grammatical inductor to appear, like in example (5b), where the conditional conjunction *si* forces the subjunctive in the verb *estar*, and the second being the one that does not need any grammatical element to appear in the sentence. This distinction mostly correlates with the subjunctive and the indicative moods, since the cases in which the subjunctive appears without depending on any inductor are highly restricted [23], and the preferred choice for most of the simple sentences is the indicative, even if in many induced contexts choosing the indicative mood is a requirement, as in example (6a).

An important characteristic of these induced contexts is that the specific mood, let it be the indicative or the subjunctive, prompted by the grammatical inductor can be an imposition, like in examples (5b) and (6b), or a choice, as in examples (7a) and (7b), and this is what is called **mood alternation**. This last case is the main feature of the experimental analysis in this thesis, and as such, it will be explained in detail next.

2.2.1 Mood Alternation

As stated above, mood alternation is the phenomenon in which a particular mood is induced within a distinct structure, usually the subjunctive in nominal complements, but this induced mood is *optional*, that is, speakers use the structure with either the indicative or the subjunctive mood, as in the examples below which were firstly introduced in Chapter 1. As was already explained, there we can see that the two sentences are almost identical, even the translations into English are the same, the only visible difference lies in the morphology of the embedded verb. *Tener* (to have) is in the indicative mood in (7a), and in the subjunctive mood in (7b). What causes this phenomenon? How is it interpreted by speakers?

(7) a. El presidente no dijo que el
the.M.SG president.M.SG not say.PST.PFV.IND.3SG that the.M.SG
país **tenía** problemas económicos.
country.M.SG **have.pst.ipfv.ind.3sg** problem.M.PL economic.M.PL
"The president didn't say that the country **had** economic problems."

b. El presidente no dijo que el
the.M.SG president.M.SG not say.PST.PFV.IND.3SG that the.M.SG
país **tuviera** problemas económicos.
country.M.SG **have.pst.ipfv.sbjv.3sg** problem.M.PL economic.M.PL

"The president didn't say that the country had economic problems."

Contrary to what we have seen in the previous section about the overall distinction between indicative and subjunctive, in mood alternation the difference is quite clear. As stated in Real Academia Española [23], Mejías-Bikandi [26] and Falk and Martin [27], the indicative is used when the speaker chooses to present the event as new information, and the subjunctive mood is used when the information is already part of the common ground. Thus, in example (7a), the fact that *the country has economic problems* is presented as something new to the speaker; and in (7b), the fact that *the country has economic problems* or that *the country doesn't have economic problems* is considered to be known by the speaker.

An important work on mood alternation is the study of Faulkner [12] from where, as it has already been shown through the references, some theoretical and practical ideas about mood alternation and its study were drawn. Her goals were to see if there are non-standard cases of mood alternation in verbal complements and if this acceptability is dialectal, to understand if the informativeness of the complements affects this acceptability, and if so, to see in which cases this happens. To fulfill these goals she collected acceptability judgments on 128 sentences, with and without context, of speakers from different Spanish varieties. For all the sentences the induced mood is the subjunctive and the alternative is the indicative.

The main idea from the results of this study is that mood alternation in verbal complements is a complex phenomenon in which different factors come into play. First of all, the type of complement, or in other words, the type of matrix verb, considerably affects the acceptability judgment. For example, desideratives don't accept the indicative in their complements, but negated epistemics do. Secondly, mood alternation seems to depend on the presence or absence of context and on whether it is informative or not. Thirdly, she shows some dialectal variations which we will neither discuss nor forget here. Lastly, it should be noted that, despite the thorough analysis she presents, not all results could be explained, which shows that further research into mood alternation is needed. Next, the specific mood alternation phenomena analyzed in this thesis are presented.

The first type is the negation of the matrix verb, shown in (7a) and (7b). As stated in Real Academia Española [23], negation can induce the subjunctive in the verb of the

nominal complement. Specifically, they say that the adverb *no* (no) induces the subjunctive mood in the complements of verbs of speech, perception, thought, and belief; and they give examples of mood alternation for verbs of perception. Given the correlation of indicative and subjunctive with new and old information, it is hypothesized that for sentence (7a) the speaker considers the embedded event *el país tenía problemas económicos* more certain or likely to be true than in (7b). Though, as it will be seen in Section 3.6, this effect depends upon the veridicality of the matrix verb. That is, whether the matrix yields the embedded predicate an entailment, a contradiction, or a neutral predicate.

(8) a. Quizás **dijo la verdad**.
maybe **tell.pst.pfv.ind.3sg** the.F.SG truth.F.SG
"Maybe s/he **told** the truth."

b. Quizás **dijera** la verdad.
maybe **tell.pst.pfv.sbjv.3sg** the.F.SG truth.F.SG
"Maybe s/he **told** the truth."

The second type of mood alternation phenomena is that of adverbs of doubt and possibility, exemplified in (8a) to (9b). As stated in Real Academia Española [23], adverbs of doubt and possibility such as quizás (maybe) or *probablemente* (probably), induce indicative or subjunctive within their sentences, but the subjunctive mood appears only when the adverb precedes the verb and there is no pause between them, as in examples (8a) and (8b). To my knowledge, research on this topic is not abundant, thus I was not able to define what exactly that *pause* means, that is, if it is literally a pause like the one indicated by *,* and other elements can be in between, such as the subject as in examples (9a) and (9b); or if it means that the adverb must inmediately precede the verb and that no element can be in between, as in examples (8a) and (8b). For the pilot study, the former option was assumed to be true, but knowing that it could *easily* be denied, is the only experiment where this phenomenon is analyzed. As to the veridicality effect, the expectations are similar to the negation case, with the only exception being that here there is no embedded event, and instead were talking about the main event, *(el presidente) dijo la verdad* (the president told the truth). Next, the second feature analyzed in this thesis is presented.

(9) a. Quizás el presidente **dijo** la verdad.
maybe the.M.SG president.M.SG **tell.pst.pfv.ind.3sg** the.F.SG truth.F.SG
"Maybe the president **told** the truth."

b. Quizás el presidente **dijera** la verdad.
maybe the.M.SG president.M.SG **tell.pst.pfv.sbjv.3sg** the.F.SG truth.F.SG
"Maybe the president **told** the truth."

2.3 Specificity

If we consider the dialog below we can see a very simple interaction in which A informs B about the location of a specific book. But that is not everything that is being communicated. If we look at the utterance that A produces, we can see that aside from the location of an entity, A is referring to two entities: a specific book and a specific table. But what does referring to something mean?

(10) A. *The book is on the table.*
B. *Ok. Thank you.*

To **refer** to something is to point, pick up, or call upon an entity or set of entities [21] in the mind of the speaker [11]. In other words, when a speaker refers, she presupposes the existence of an entity or set of entities and connects a linguistic expression to it [28]. When the listener hears such an expression, he tries to make that same connection by corroborating its existence in the physical context, linguistic context, and/or his memory. This is why, the study of referential expressions is intrinsically connected to existence [21], to the study of existential presuppositions; to the point that, as stated in García Murga [28], the study of existential presuppositions must be supported by the study of reference. Further clarifications about this relation can be seen in Lyons [21], García Murga [28] and Herrasti et al. [29].

Above we have mentioned that the book and table to which A refers are a specific book and a specific table. At first glance, this can be simply understood as a particular book and a particular table, but as Caudet [11] thoroughly explains, the property of being specific has been applied to different concepts, which can lead to some confusion. Thus she puts together the different views on specificity and summarizes them into the following six criteria, criteria that can be used to classify referential expressions:

1. Existence of the referent in the extra-linguistic reality.
2. Identifiability of the referent in the extra-linguistic reality.
3. Extension indicated by the referential expression.
4. Existence of the referent in the discourse universe.
5. Identifiability of the referent in the discourse universe.

6. The expression points to a set of referents pragmatically delimited or chooses a subset of them.

To illustrate the first criterion a good example would be the distinction between "The money I have." and "The money I will have when I'm rich". In the first case "the money" points to real money that the speaker has and that can be counted. Contrary to this, in the second case "the money" is not one that the speaker actually has or that we can feel and count, but rather one that at most exists only in the mind of the speaker. Although this distinction of existing or not in the extra-linguistic reality was shown with a material example, it should be noted that applies to all kinds of entities.

As to the second criterion, it shows the distinction between pointing to a known and an unknown entity, but it is a pragmatic distinction, it does not affect the existence of the entity. So "my daughter's boyfriend" is still unknown by the speaker in "I will meet my daughter's boyfriend for the first time" and known in "My daughter's boyfriend is again coming for dinner today", but in both cases, he exists.

The criterion of extension deals with quantity. Specifically with the number of entities or the amount of matter the referential expression points to. This criterion is exemplified in sentences like "He doesn't have any kids" vs. "He has six kids" or "Most of the flour is on the floor" vs "Almost nothing of the flour is on the floor".

The fourth criterion is the one intrinsically connected to coreference and anaphora, so that instead of being built with morphological or syntactical features, it is constructed throughout the discourse and it distinguishes between the entities that have already been introduced and those which are just being presented like "the eldest daughter" in "My daughters have just arrived. The eldest one is crying" vs. "My eldest daughter is crying".

Regarding the identifiability of the referent in the discourse universe, it involves the accessibility of the referent, which is shown, for example, in the amount and type of information used in the referral expression. So that when referring to Olaf Scholz, the current German chancellor, we could use the expression "the German chancellor", "the chancellor", or directly "Olaf Scholz".
The last criterion applies to the distinction between "The students will go on the school trip" and "The students who have passed all of their courses will go on the school trip". If this utterance was to be uttered by, for example, a 6^{th} grade teacher, in the first case he would be referring to all of his students, whereas in the second one, he would be

pointing to only of a subset of his students.

From these six criteria, here we consider **specificity** in the sense of the fifth criterion, therefore when manipulating the specificity of a nominal phrase, we are influencing the accessibility of the referent in the discourse universe is what is being analyzed. In particular, here we manipulate it by modifying the amount and type of information given in the subject of the premise.

Regarding the type of information, two manipulations are performed: individual vs. collective nouns, and common vs. proper nouns. With the first one, we are manipulating the number of individual entities to which we are referring. As it was mentioned in Chapter 1, by using an individual noun (in singular) like *president*, we point to one single entity, whereas by using a collective noun like *government*, we point instead to a set of entities; thus when trying to identify the referent in the discourse universe, the set of possible referents the listener considers is different.

As to the second manipulation, common vs. proper nouns, we alternate who identifies the referent. When using a common noun like *president*, the speaker only identifies the set of possible entities to which our referent belongs, therefore leaving to the listener the final determination of the particular referent. Contrary to this, when using a proper noun, the speaker *usually*[2] points directly to an entity or set of entities. Therefore, by distinguishing between common and proper nouns we have two different ways of identifying the referent: identification by the listener and identification by the speaker.

Concerning the amount of information, only one manipulation is considered: is the common noun in the subject modified only by a determiner as in *the president*, or are there any other complements that modify the noun as in *the president of the government*? By performing this manipulation, which here we denominate mixed, we are influencing the retrieval of the referent in a similar way as to the common vs. proper noun distinction since we are modifying the number of possible referents the listener considers. When we have a mixed noun phrase, this number is smaller than with an almost bare noun phrase, but bigger than with a proper noun, thus being a sort of middle ground between common and proper nouns and resulting in the following scale of number of possible referents: common > mixed > proper.

[2]For some exceptions on the common use of proper nouns see Caudet [11].

Regarding the veridicality of these specificity manipulations, it is difficult to define an exact hypothesis since, to my knowledge, there is no work in the matter and thus expectations are surrounded with more uncertainty than mood alternation. Overall it is hypothesized that referring to a set of entities instead of a single one will increase certainty since collective nouns often refer to institutions and organizations for which there is a certain degree of trust, whereas for individual agents subjectivity is assumed to play a greater role and therefore increase uncertainty to even a negative point. Therefore, with a subject like "The police arrested a suspect" a more positive factuality judgment about the fact that the suspect was arrested than in "The police officer arrested a suspect". A similar assumption is applied to the common > mixed > proper scale, with common nouns having a more positive veridicality effect than proper nouns, again due to the subjectivity associated with them.

Chapter 3

Pilot Study

3.1 Design

The goal of this pilot study is three-folded. The first one is crowdsourcing annotations on premise-hypothesis pairs to answer the four research questions introduced in Chapter 1 and which are repeated below:

RQ1.- In a complex sentence like examples (7a) and (7b), how does the mood alternation of the embedded verb that occurs due to the negation of the main or matrix verb affect the factuality value of the embedded event?

RQ2.- In a simple sentence like examples (9a) and (9b), how does the mood alternation caused by an adverb of doubt or possibility affect the factuality judgment of the event?

RQ3.- How does an individual subject, like the one in example (9a), affect the factuality judgment of the event?

RQ4.- How does a subject that refers to a collective entity like *the government*, affect the factuality judgment of the event?

These questions yield the four experimental conditions: negation (RQ1.-) and with examples (11) to (13) , adverb or possibility (RQ2.-) and exemplified in (14) to (16), individual (RQ3.-) and shown in (11) to (13), and collective (RQ4.-) presented in (14) to (16). Furthermore, given the novelty of these conditions and to ensure experimental soundness, the second and third goals of the study are to ensure the correctness of the experimental design and procedure and to explore different possibilities of statistical analysis, so that in the final study the data is properly analyzed and represented.

(11) El hijo oyó que le estaban
the.M.SG son.M.SG hear.PST.PFV.IND.3SG that him be.PST.IPFV.IND.3PL
llamando.
call.GER
The son heard that they were calling him.

(12) El hijo de la actriz no oyó que le
the.M.SG son.M.SG of the.F.SG actress.F.SG not hear.PST.PFV.IND.3SG that him
estaban llamando.
be.PST.IPFV.IND.3PL call.GER
The actress' son didn't hear that they were calling him.

(13) El hijo de Nicole Kidman no oyó que le
the.M.SG son.M.SG of Nicole Kidman not hear.PST.PFV.IND.3SG that him
estuvieran llamando.
be.PST.IPFV.SBJV.3PL call.GER
Nicole Kidman's son didn't hear that they were calling him.

If we take the four conditions, sort them into two groups, mood alternation for negation and adverb, specificity for individual and collective; and cross them, we obtain a 2x2 design, where each condition is divided into 3 subconditions or categories. The negation and adverb conditions are divided into three categories: baseline (B) shown in examples (11) and (14), where there is no negation of the matrix verb or presence of a doubt or possibility adverb; indicative (I), portrayed in examples (12) and (15), where we have a mood induced adverb and the verb is in indicative mood; and subjunctive (S), exemplified in (13) and (16), where we have a mood induced adverb and the verb is in the subjunctive mood.

The specificity conditions are also divided into three categories: common (C), where the subject is a common noun with the determiner as its only complement as in examples (11) and (14); mixed (M), where the subject is a common noun but with a determiner and another complement like examples (12) and (15); and proper (P), where the subject is directly a proper noun like example (16) or there is a relation to a proper noun, like in example (13).

(14) El sindicato tenía otra idea.
the.M.SG union.M.SG have.PST.PFV.IND.3SG another.F.SG/a.different idea.F.SG
The union had another/a different idea.

(15) Quizás el sindicato de la empresa
maybe the.M.SG union.M.SG from the.F.SG company.F.SG
tenía otra idea.
have.PST.PFV.IND.3SG another.F.SG/a.different idea.F.SG

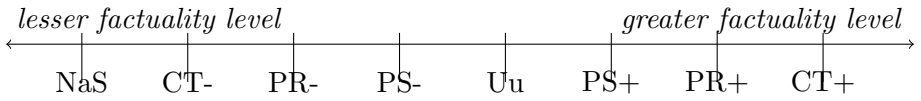

FIGURE 3.1: Ordered representation of the labels used for annotating the corpus. Each label stands for: certainly yes (CT+), probably yes (PR+), possibly yes (PS+), unknown or uncommitted (Uu), possibly not (PS-), probably not (PR-), certainly not (CT-), not a sentence (NaS).

Maybe the company's union had another/a different idea.

(16) Quizás Comisiones Obreras tuviera otra
maybe Comisiones Obreras have.PST.PFV.SBJV.3SG another.F.SG/a.different idea.
idea.F.SG

Maybe Comisiones Obreras had another/a different idea.

As to the set of labels used to annotate each pair, as in de Marneffe et al. [3] the labels from Saurí and Pustejovsky [2] minus *certain but unknown output* (CTu) were used. But since, as mentioned in the previous chapter, the acceptability of sentences like (9a), where we have a nominal phrase between the adverb of possibility and the verb, was not certain, another label, *not a sentence* (NaS), was added resulting in the following set of 8 labels, represented for simplicity as a one-dimension scale in Figure 3.1: certainly yes (CT+), probably yes (PR+), possibly yes (PS+), unknown or uncommitted (Uu), possibly not (PS-), probably not (PR-), certainly not (CT-), not a sentence (NaS).

Data Generation within Experimental Design I					
			SPECIFICITY		
			Individual		
			C	M	P
MOOD	Negation	B	S1	V1	V1
		I	V1	V1	V1
		S	V1	V1	V1
	Possibility	B	S3	V3	V3
		I	V3	V3	V3
		S	V3	V3	V3*

TABLE 3.1: Experimental conditions and corpus generation first part. Each cell is for 9 pairs. **S** followed by a number stands for set of seeds, **V** stands for variants, with the number indicating the corresponding set of seeds. ***** indicates where the 2 mistaken pairs are. **B, I, S** stands for *baseline, indicative, subjunctive*, and **C, M, P** stands for *common, mixed, proper.*

3.2 Dataset

As to the dataset to be annotated, it was constructed by first writing 9 seed premises for each of the four baseline-common combinations, and from them, the hypotheses were extracted to form the pairs. Then these seeds were changed across conditions, yielding a total of 324 premise-hypothesis pairs[1]. Tables 3.1 and 3.2 show this generation process together with the experimental design.

Data Generation within Experimental Design II					
			SPECIFICITY		
			Collective		
			C	M	P
MOOD	Negation	B	S2	V2	V2
		I	V2	V2	V2
		S	V2	V2	V2
	Possibility	B	S4	V4	V4
		I	V4	V4	V4
		S	V4	V4	V4

TABLE 3.2: Experimental conditions and corpus generation second part. Each cell is for 9 pairs. **S** followed by a number stands for set of seeds, **V** stands for variants, with the number indicating the corresponding set of seeds. ***** indicates where the 2 mistaken pairs are. **B, I, S** stands for *baseline, indicative, subjunctive*, and **C, M, P** stands for *common, mixed, proper*.

3.3 Predictions

Now that the different variables considered in the study have been explained, we can try to envision how they will affect the annotations. As we can see in tables 3.3 and 3.4, the highest factuality labels are predicted for the baseline category, since in the absence of any of the mood-inducing conditions, no other factuality modifier is expected, for the exception of the specificity categories. When combined with them, the factuality level of the baseline pairs might change, but only within the range specified in each case. This is because by modifying the amount of information (mixed category) or the type of information (proper category), it is expected that world knowledge, and thus uncertainty, plays a greater role in assigning the label.

As to the other two mood categories, indicate and subjunctive, a decrease in the factuality level relative to the baseline is to be expected, with the indicative having a higher

[1]Upon reviewing the pairs after the experiment was run, it was noticed that there was a mistake on 2 of them, thus yielding them invalid. The whole two seeds were taken out for analysis, yielding a total of 306 analyzed pairs.

Label Predictions within Experimental Design I					
			SPECIFICITY		
			Individual		
			C	M	P
MOOD	Negation	Baseline	CT+	CT+-PR+	CT+-PS+
		Indicative	PR+	PR+-PS+	PR+-Uu
		Subjunctive	PS-	PS+-PS-	PS+-PS-
	Possibility	Baseline	CT+	CT+-PR+	CT+-PS+
		Indicative	PR+	PR+-PS+	PR+-Uu
		Subjunctive	PS+	PS+-Uu	PS+-Uu

TABLE 3.3: First part of rough approximation of label predictions for each combination of experimental categories. **C, M, P** stands for *common, mixed, proper*, and - indicates a range of labels when not indicating the quality of a specific label.

degree of factuality than the subjunctive. Since by choosing the indicative mood the speaker presents the event, our hypothesis, as new information, she limits the possibilities of the listener belying its factuality. Contrary to this, by using the subjunctive, the speaker assumes that the factuality of the event is already known and agreed upon, thus increasing the possibility of it being denied, or in other words, she allows more uncertainty. These predictions for the indicative and subjunctive categories roughly hold for both mood alternation conditions, with the only difference being that for the negation condition, the factuality level for the subjunctive category is predicted to be lower than for the same category in the possibility condition, reaching even the level of the negative labels. This is due to the above-mentioned fact that only in this case the negation marker scopes over the embedded event, thus reducing the factuality of the event more strongly than a mere adverb of doubt or possibility.

Regarding their variability when crossed with different specificity categories, it is expected that labels within the indicative category will decrease similarly to labels within the baseline category; but for the subjunctive category, barely any differences are expected since values in the subjunctive category are already within the range of high uncertainty. Furthermore, it is expected that events under collective subjects will have a greater factuality than those under individual subjects, but not as big as to be easily shown with specific labels.

Lastly, it should be noted that for most of the cases, the labels predicted are within the positive range of the scale. In other words, the distribution of the labels is expected to be negatively skewed.

Label Predictions within Experimental Design II					
			SPECIFICITY		
			Collective		
			C	M	P
MOOD	Negation	Baseline	CT+	CT+-PR+	CT+-PS+
		Indicative	PR+	PR+-PS+	PR+-Uu
		Subjunctive	PS-	PS+-PS-	PS+-PS-
	Possibility	Baseline	CT+	CT+-PR+	CT+-PS+
		Indicative	PR+	PR+-PS+	PR+-Uu
		Subjunctive	PS+	PS+-Uu	PS+-Uu

TABLE 3.4: Second part of rough approximation of label predictions for each combination of experimental categories. **C, M, P** stands for *common, mixed, proper*, and - indicates a range of labels when not indicating the quality of a specific label.

3.4 Procedure

The annotations were collected through Google Forms. The first page of the form asked them to choose their variation of Spanish (European or American)[2], and the second one showed the instructions for the task, together with an example. It should be noted that these instructions might have been too simple, and it was decided that they should be elaborated more carefully for the main study.

After the second page, the pairs to be labeled were presented on 9 pages. Following de Marneffe et al. [3], the task consisted of questions where the labels were presented as answers in a single-choice list. de Marneffe et al. [3] presented the labels ordered, but, as seen in Figure 3.2, in this study labels were by mistake not completely ordered.

Another important feature of this experimental procedure is that, since cognitive overload and an overly time-consuming task were to be prevented, pairs were divided into 9 batches, thus having a total of 9 forms with 36 pairs each. Initially, only 3 annotators were to annotate each pair, but as the task progressed, the lack of agreement was noticed, and therefore two more annotators were added to each batch. That is, pairs were annotated by 5 raters, and each of them annotated 36 pairs.

Concerning the annotators, as mentioned in Chapter 1, they were chosen among family and friends whose mother tongue is Spanish and who do not have higher linguistic education, in other words, linguistically naive speakers. This last feature was important

[2]Given that the number of annotators for each group was completely different, this variable was not taken into account for the analysis.

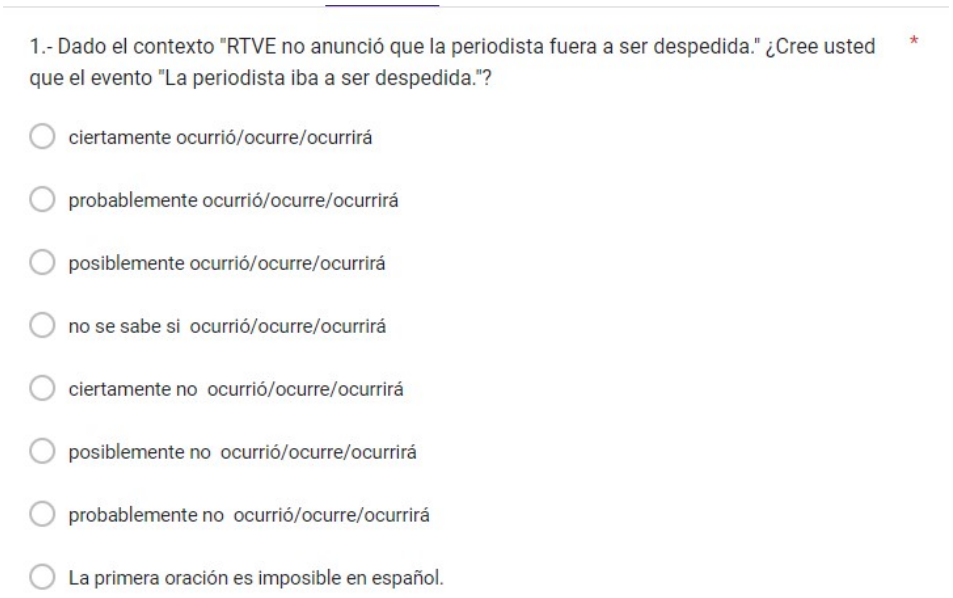

FIGURE 3.2: Interface for the pilot study. The order here is: CT+, PR+, PS+, Uu, CT-, PS-, PR-, NaS. It should have been: CT+, PR+, PS+, Uu, PS-, PR-, CT-, NaS

to prevent annotations from having a lexical approach, but, some of the annotators acknowledged afterward that, since they attempted to find the *correct* answer, they had to use their linguistic knowledge from their basic education instead of their *raw* linguistic instincts; hence this was acknowledged in the instructions for the final study.

Now that we have explained the study goals, design, predictions, and procedure, we can present the statistical analysis of the annotations collected.

3.5 Results

3.5.1 Overview of the Annotations

To get a first understanding of the annotations, different plots were drawn and some very basic statistics were computed. These computations are shown in Table 3.5 and the two most relevant plots are shown in figures 3.3 and 3.4. Probably the two most important things to notice are that, as predicted, in Figure 3.3, where we have the overall distribution of labels, we observe that this distribution is quite negatively skewed, despite the considerable amount of pairs with a negation marker; and that the number of pairs for which agreement is reached is rather low, considerable lower than in de Marneffe et al.

Basic Statistics		
	Absolute	Percentage
Total number of analyzed pairs	306	100
Total number of analyzed annotations	1530	100
Number of pairs without a label	121	39.542
Number of pairs with a label	185	60.458
Agreement equal to 3	112	36.601
Agreement greater than 3	73	23.856
Maximal agreement	18	24.657
Number of pairs with at least one NaS	32	10.457
Number of times NaS label was used	37	2.418
Number of baseline pairs in which NaS was used	8	25
Number of pairs with NaS labeled as NaS	0	0
Number of problematic seeds	8	23.530

TABLE 3.5: Some basic counts of the annotations. Maximal agreement means that the 5 raters agreed on one label. A problematic seed has 5 or more of its variants without a label.

[3], where the percentage of pairs without a label was 22.118% since the percentage of pairs for which it was possible to compute an aggregated label is 60.458. In that same plot, we see that the differences between the negative labels are rather small and that the Uu and PR+ labels are used with almost identical frequency.

In Figure 3.4, where the distribution of labels per batch is shown, there is further evidence of a considerable lack of agreement, since several differences between the label distributions for each batch appear, contrary to what was expected. For example, we see that batch B clearly uses the label CT+ more often than batch E, or that batches H and I seem to use the label NaS more often than the others.

Another important observation is the use of the label NaS. This label, mainly meant to be an insurance for the possibility condition, seems to be overused. This is evident in the fact that although NaS was used at least once in 10.457% of the pairs, no pair was labeled as not acceptable; furthermore, 8 pairs in which this label was used, were baseline pairs, that is, pairs without any expected difficulty. Therefore the role of this label must be reviewed.

To sum up, in this section, we have seen that, as predicted, the distribution of the labels is negatively skewed, with most of the counts being in the positive range of the labels' scale. We have also shown that the NaS label might have been misused in the annotations. Moreover, some indications of a considerable amount of disagreement in

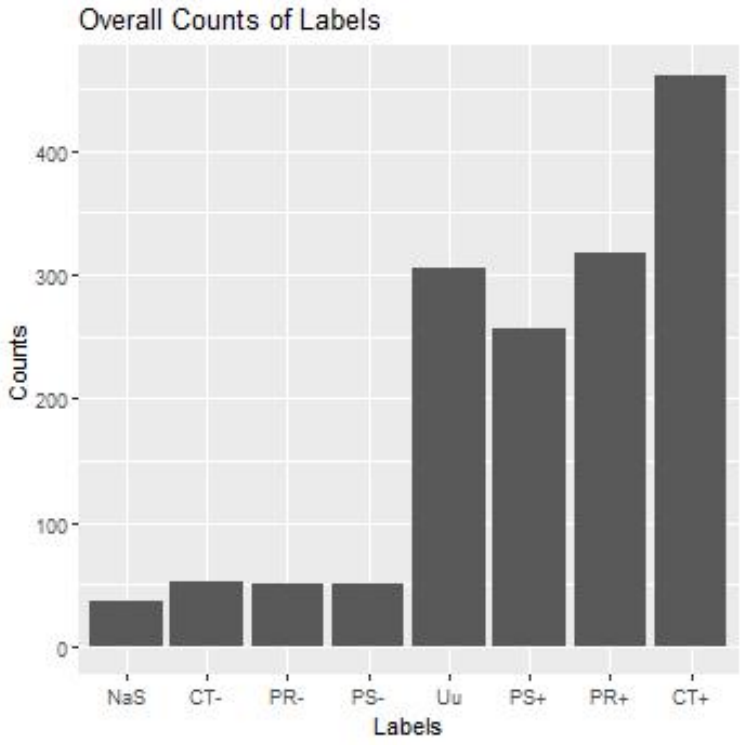

FIGURE 3.3: Overall distribution of the proportion of labels used by annotators.

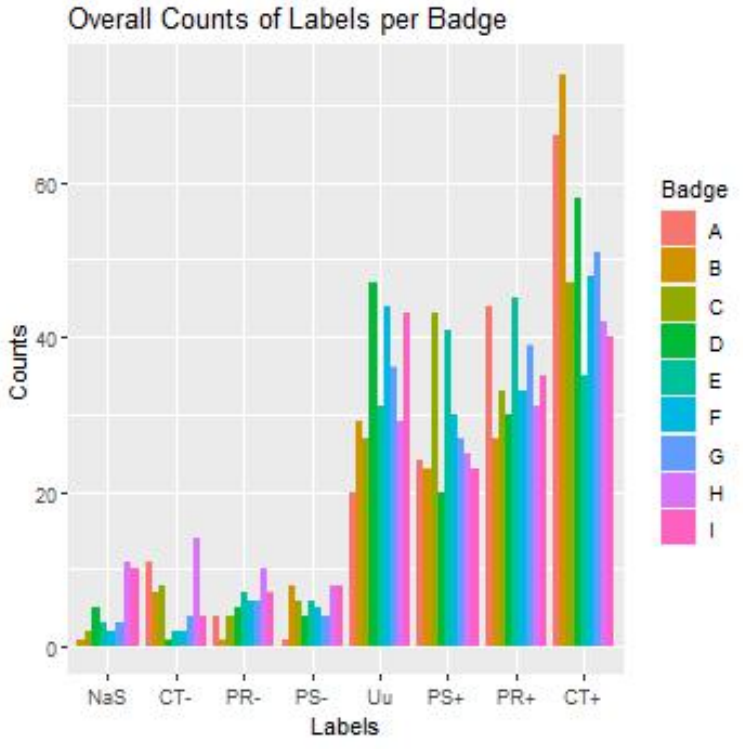

FIGURE 3.4: Distribution of the proportion of labels used by annotators grouped by batch.

the annotations have been presented: the number of pairs in which at least 3 annotators agreed upon a label is lower than in the case of de Marneffe et al. [3] and the labels' distributions for each batch have some clear differences. In the next section, we will see the significance of this disagreement by measuring inter-annotator agreement.

3.5.2 Inter-Annotator Agreement Scores

Finding the score that correctly computes the inter-annotator agreement was a more complicated task than expected. Thus, instead of simply presenting one κ value to represent the inter-annotator agreement for the whole corpus, values for different scores are presented. All inter-annotator agreement scores computed for this study can be seen in the appendix.

Initially, following the work of de Marneffe et al. [3], Fleiss' kappa [30] was computed, which resulted in a value of $\kappa^f = 0.160$, which means slight agreement according to the labels presented in Shrout [31], but since, as de Marneffe et al. [3] already remarks, this score does not consider the order between the labels shown in Figure 3.1, thus yielding an inaccurate value for this case, other possibilities had to be considered.

The main alternative is to use weighted kappa, a score that extends Cohen's kappa [32] to observations made using ordinal labels. For the case presented here, we would specifically need to extend the definition to the case of multiple raters and ordinal labels, a task that is not without problems [33, 34] and thus it has different solutions. What all the different implementations have in common is defining κ as a relation between the weighted proportion of observed agreement p_{aw}, and the weighted proportion of chance agreement $p_{a|c_w}$ [35].

Here, initially, the definitions presented in Warrens [36] and Warrens [37] were chosen, but after encountering the work of Vanacore and Pellegrino [35], this decision changed. Vanacore and Pellegrino [35] presents a theoretical analysis of four weighted kappas which differ in their definition of $p_{a|c_w}$, Fleiss [30] (κ_w^f), Conger [38] (κ_w^c), Brennan and Prediger [39] (s_w^*), and Gwet [40] (AC_2); as well as a comparison of their behaviors in paradoxical environments. Their results showed that both κ_w^f and κ_w^c were the most affected by these environments. So, given these results, the fact that κ_w^f and κ_w^c are, to my knowledge, more commonly used; and that, as seen in Figure 3.3, the distribution of our labels is skewed, or in other words, paradoxical; three different agreement scores

Agreement Scores for the Whole Dataset	
Name	Value
Fleiss' κ	0.160
Fleiss' κ with linear weights	0.177
Conger's κ with linear weights	0.176
Gwet's AC_2 with linear weights	0.484

TABLE 3.6: Inter-annotator agreement scores computed for the whole dataset.

were computed, κ_w^f, κ_w^c, and AC_2, with the software package R, specifically the IrrCAC library [41]. Results are shown in Table 4.6.

As we can see, all weighted scores show an improvement concerning the unweighted Fleiss' κ, but only AC_2 clearly distinguishes itself from the kappa computed first. The other two weighted kappas are not just within the same range of sligh agreement $(0.11 - 0.40)$, but there are also barely distinguishable from the unweighted kappa. This could be interpreted as more evidence of a very low agreement between annotators, but, given the already mentioned work of Vanacore and Pellegrino [35], that in 60.458% of the pairs, at least 3 annotators agreed upon a label, and that the AC_2 has a value of 0.484, which is clearly within the range of fair agreement $(0.41 - 0.60)$; it appears rather that these results can be taken as more evidence of the sensibility of κ_w^f and κ_w^c to paradoxical behavior. Thus it seems appropriate to consider $AC_2 = 0.484$ as the inter-annotator agreement score that represents our corpus, but to strengthen this decision, the three weighted scores are used in any other computations presented here. Now we will briefly explore agreement within the experimental conditions.

3.5.2.1 Inter-Annotator Agreement for each Experimental Condition

Table 3.7 shows the value of the 3 inter-annotator agreement scores computed for the subsets of the corpus that correspond to each of the experimental conditions. Again we see what we saw above, Gwet's AC_2 is clearly greater in each case, and Fleiss' and Conger's κ are roughly equal, thus strengthening the decisions to use Gwet's AC_2 score. More importantly, based on these scores we can say that the difference between the specifcity conditions is minimal, and therefore they are likely to be quite irrelevant as veridicality categories. Contrary to this, the difference between the mood alternation conditions is quite big, suggesting that they are suitable as veridicality categories. We will see afterward if these ideas hold or not. Next some simple computations done to explore the validity of the set of labels used are presented.

Agreement Scores for Experimental Conditions			
Subset	κ_w^f	κ_w^c	AC_2
ALL	0.177	0.179	0.484
Negation	0.094	0.097	0.388
Possibility	0.278	0.278	0.592
Individual	0.161	0.162	0.500
Collective	0.190	0.192	0.470

TABLE 3.7: Inter-annotator agreement scores computed for the subsets corresponding to each of the experimental conditions. The value for the whole dataset is given as a reference.

3.5.3 Label Space Reductions

To verify the validity of the label set, or in other words, to check if the low agreement score could be partially blamed on an inadequate label set, different label space reductions were performed, and the three weighted scores were computed for each of these reductions. The reductions are of two types: removal of any pair which was annotated with the label NaS, and merging of pairs of labels. Merging is understood as the replacement of any instance of the first element of the pair with the second element, with the corresponding quality signs when required.

The results, shown in Table 3.8, display very interesting tendencies. First of all, the behaviors of the scores with the whole label set still hold: κ_w^f and κ_w^c are quite similar and within the range of slight agreement, and AC_2 is considerably higher and, for most of the cases, within the range of fair agreement. Secondly, we see that the scores are not always affected in the same way by the different reductions. Probably the best two examples are the *No NaS* and the *PR with CT, PS with Uu* reductions. In the first case, we see that removing all pairs in which the NaS label is used benefits both κ_w^f and κ_w^c, but it has a detrimental effect on AC_2. In the second case, we witness the opposite effect happening. When reducing to the typical labels {*yes, unknown, no*} (CT+, Uu, CT-), with or without NaS, it seems that it greatly benefits AC_2, but that it diminishes both κ_w^f and κ_w^c. Thus, it appears that making radical decisions about the label set might be unwise. Nevertheless, it was considered that removing PS+ and PS- could be beneficial for the final study since scores increased in every case they were removed.

Next, the results of fitting a regression model to the annotations are presented.

Agreement Scores for Different Label Reductions			
Reduction	κ_w^f	κ_w^c	AC_2
ALL	0.177	0.179	0.484
No NaS	0.190	0.191	0.469
PS with PR	0.186	0.187	0.572
PS with PR, no NaS	0.204	0.205	0.537
PR with CT, PS with Uu	0.153	0.155	0.635
PR with CT, PS with Uu, no NaS	0.162	0.164	0.518
PS with Uu	0.204	0.205	0.590
PS with Uu, no NaS	0.224	0.225	0.550

TABLE 3.8: Inter-annotator agreement scores computed for the different label space reductions. *No NaS* means that all pairs in which any of the annotators used the label NaS were removed. For any pair of labels linked by the preposition *with*, the interpretation is that the first label was replaced with the second one, with the corresponding quality signs when required.

3.5.4 Model Fitting

To understand how the different variables defined in Section 3.1 influenced the annotations, a cumulative link mixed model, was fitted with a logit link by using the R software, specifically the ordinal package [42]. As predictors, mood conditions were crossed with specificity conditions and mood categories, specificity conditions were further crossed with specificity categories, and mood categories were also crossed with specificity categories. Items and annotators were set as random intercepts, and the labels as the response variable. Here only the most relevant results are presented, all details are in the appendix.

The most important observation is that most of the effects defined do not have a significant effect, only the mood condition adverb ($p = 2.31 \times 10^{-14}$) on its own, and crossed with the different mood categories ($p < 2 \times 10^{-16}$ for both the indicative and subjunctive category) have significant effects, which is consistent with the results in Section 3.5.2.1. This means that the veridicality effect of the possibility condition is significantly different from the negation condition and that within the former, the actual mood alternation categories are significantly different from the baseline, which means that at least with this condition the baseline is clearly established. As to the other predictors, there is one that is close to being significant, that of the crossed effect of the mood condition adverb with the specificity condition collective ($p = 0.054$), telling us that for the adverb condition, the specificity conditions might have a significant effect if the experimental design and procedure are improved.

Model Threshold Coefficients			
Threshold	Estimate	Std. Error	z value
NaS—CT-	−4.254	0.326	−13.062
CT-—PR-	−3.272	0.298	−10.962
PR-—PS-	−2.756	0.291	−9.480
PS-—Uu	−2.371	0.286	−8.277
Uu—PS+	−0.904	0.278	−3.249
PS+—PR+	0.006	0.277	0.023
PR+—CT+	1.233	0.279	4.415

TABLE 3.9: Threshold coefficients resulted from fixing a cumulative link mixed model to the whole dataset with different combinations of mood conditions, mood categories, specificity conditions, and specificity categories; and random intercepts for raters and premise-hypothesis pairs.

The model threshold coefficients, also known as cut-points or intercepts [42] and shown in Table 4.11, represent the division of the *true* underlying factuality value for a premise-hypothesis pair. Thus they can be interpreted as the width of each category in the label set, except our two extremes of the scale. Considering this and that, aside from the PR- and PS- labels, the space for each label is roughly similar with values ranging within $(0.5-1.2)$, we can say that the labels are all informative. Given this and the fact that the computations presented in the previous section were more an exploration than a proper statistical test, the proposal to remove the PS+ and PS- labels was not considered.

Lastly, it should be noted that the variance for both random variables is almost, 0.512 for pairs and 0.505 for annotators, suggesting that there is more information to be extracted from this data.

To sum up, from implementing a cumulative link mixed model it was learned that mood conditions and mood categories are informative to the experimental design, although not in all settings, and that the specificity conditions could have a significant effect when combined with the adverb condition. Hence it appeared reasonable to keep these three categories of variables for the final study, but not the specificity categories. Furthermore, given the threshold coefficients obtained, the proposal to remove the PS+ and PS- labels was no longer considered. Next, we will proceed to the final section of the analysis of results where the annotations collected will be compared to the predictions.

3.5.5 Comparison with Predictions

Once we have obtained some numerical features of our annotations, we can dive into the data to examine whether the predictions were fulfilled, and what linguistic characteristics favored or hindered agreement. To do so, the labels assigned to each combination of the experimental conditions and categories presented in Table 3.10 and Table 3.11 are examined.

Assigned Labels within Experimental Design I					
			SPECIFICITY		
			Individual		
			C	M	P
MOOD	Negation	B	CT+/PR+/Uu	CT+/PS-	PR+/Uu
		I	CT+	CT+	-
		S	?	CT+/Uu	-
	Possibility	B	CT+	CT+	CT+
		I	PR+/Uu	?	PR+/PS+
		S	PR+	-	?

TABLE 3.10: Assigned labels for the first half of each combination of experimental categories. **B, I, S** stands for *baseline, indicative, subjunctive*, and **C, M, P** stands for *common, mixed, proper* The presence of a label/s or **?** means that there were at least 5 pairs in which 3 raters agreed upon a label, the opposite situation is indicated with **-**. **?** means there was no majority vote for any label among the different pairs. **/** means that the votes were equally divided.

There are two main remarks to make about these tables. First of all, for the indicative category there is a higher factuality level than expected, particularly when crossed with the negation condition and to the point that, in some cases, the label assigned to the indicative pair has a higher factuality label than for its correspondent baseline pair, although the distance between these labels is, in more cases, minimal. Furthermore, heeding the considerable lack of agreement and the consequential lack of pairs with a label, it seems that this difference between indicative and baseline pairs is consistent across all variants of the respective seed. Thus we can say that this difference is due to some lexical characteristics. Further analysis into this phenomenon was left for the final study.

The second important observation is two-folded and concerns the labels assigned to the baseline pairs in the negation condition. First of all, we see a higher disagreement than expected, indicated both by the lack of variants to set a label for the combination and by the label split, in other words, more than one label is assigned, for half these combinations. Second of all, we see, in some cases, a lower factuality level than expected, to the point of uncertainty (Uu label). Determining the cause of these phenomena is not straightforward, but there is one helpful fact. Upon reviewing the data, it was noticed

that one feature was not taken into consideration when defining the predictions: the matrix verbs.

Assigned Labels within Experimental Design II					
			SPECIFICITY		
			Collective		
			C	M	P
MOOD	Negation	B	CT+	-	-
		I	CT+	-	CT+
		S	-	-	-
	Possibility	B	CT+	CT+	CT+
		I	PR+	-	?
		S	PR+/Uu	Uu	-

TABLE 3.11: Assigned labels for the second half of each combination of experimental categories. **B, I, S** stands for *baseline, indicative, subjunctive*, and **C, M, P** stands for *common, mixed, proper* The presence of a label/s or **?** means that there were at least 5 pairs in which 3 raters agreed upon a label, the opposite situation is indicated with **-**. **?** means there was no majority vote for any label among the different pairs. **/** means that the votes were equally divided.

Matrix verbs, that is, predicates that embed another predicate or event, have their veridicality value, or, in other words, they affect the factuality value of the predicate they embed. According to this effect, matrix verbs are classified into three groups: implicative, factive, and epistemic. The first group designates the matrices whose embedded events are a consequence of the whole complex event, like *manage* in *He managed to sell the house*. The second one denotes those whose embedded predicate is a fact or a precondition for the whole complex predicate, like *know* in *He knew her father had arrived*. Lastly, the group of epistemic verbs includes those whose embedded event is a possibility, like *think* in *He thinks that Peter has arrived*. An extreme case of how this distinction affects the baseline of our annotations are examples 17 and 18, whose assigned labels are CT+ and PS+ respectively.

(17) a. La Razón sabía que la noticia
La Razón know.PST.IPFV.IND.3SG that the.F.SG news.F.SG
era falsa.
be.PST.IPFV.IND.3SG fake.PTCP.M.SG
La Razón knew that the news was fake.

b. La noticia era falsa.
the.F.SG news.F.SG be.PST.IPFV.IND.3SG fake.PTCP.M.SG
The news was fake.

In the first case, the matrix verb *saber* (to know) makes our hypothesis 17b a fact, or in other words, it assigns a high factuality value from our label scale, like *certainly yes*, to the pair. This assignment can surely be modified by other factors, but it definitely serves as a basic guideline to explain its behavior concerning the label of the pair below. In this second pair, the verb *pensar* (to think), is an epistemic verb, thus yielding 18b a possibility, and therefore assigning to it a more uncertain factuality value to begin with than for 17b. Furthermore, if we look at the other baseline variants from these pairs, we see more signs of this behavior. In the case of 17, the other 2 variants have the same assigned label, CT+, whereas in the case 18, the other variants do not even have an assigned label. Given these differences, it will be prudent to consider their classification when defining the predictions and analyzing the data from the final study.

(18) a. La junta directiva del Barça
the.F.SG board.F.SG governing.F.SG of.the.M.SG Barça
pensó que la reunión con los
think.PST.PFV.IND.3SG that the.F.SG meeting.F.SG with the.M.PL
abogados había ido bien.
lawyer.M.PL have.PST.IPFV.IND.3SG go.PTCP.M.SGwell
Barça's board of directors thought that the meeting with the lawyers had gone well.

b. La reunión con los abogados había
the.F.SG meeting.F.SG with the.M.PL lawyer.M.PL have.PST.IPFV.IND.3SG
ido bien.
go.PTCP.M.SGwell
The meeting with the lawyers had gone well.

As to the differences between the mood categories, there is not enough agreement to make clear conclusions, but some remarks can be pointed out. First of all, the distinction between the baseline and the mood alternation categories in the possibility condition is quite clear. All adverbial baseline combinations have a higher factuality value than the other two categories, but the distinction between indicative and subjunctive is not so clear. In the case of the negation condition, is more difficult to establish tendencies given the greater lack of agreement, but it seems that there is a difference between the three categories, although it is not clear how it works.

Further observations can be made about tables 3.10 and 3.11. The first one is that, consistent with the model estimates, the specificity categories do not appear to affect the resulting labels, thus it might be helpful to remove them for the final study. Contrary to this, it seems that the specificity conditions affect the labels, particularly in the

negation condition and even if it is to cause more disagreement. Lastly, there are some signs of a label split as seen in the baseline of the mood condition, but it could be due to the above-mentioned issue of the matrix verb, and therefore more data and a more thorough analysis are needed.

To sum up, in this section evidence that can partially explain the unexpected values seen in the baseline category has been presented, also we have seen how there are some distinctions between the different mood conditions and categories, even if for the latter the significance of their effect is not yet clear. Likewise, there are a few signs of variations between the specificity conditions but not the specificity categories. Lastly, we have also seen the overall need for a deeper linguistic analysis of the annotated pairs, an analysis that will be presented in the corpus of the final study.

3.6 Discussion

Throughout this chapter, different results have been presented. Out of these, one of the main ones is the inter-annotator agreement, $AC_2 = 0.484$, that although higher than the initially chosen score, is still lower than previous research on veridicality, with the work of de Marneffe et al. [3] having a $\kappa^f = 0.53$ and that of Ross and Pavlick [17] having an average Spearman correlation of 0.78 for positive contexts and of 0.74 for negative contexts. , thus causes of this lower agreement need to be considered.

As already mentioned in Section 3.4, some factors in the experimental procedure could have influenced the results. First of all, there was an important mistake: the labels were not presented in order. This could have prevented annotators from internalizing the order behind the label set and thus made it more difficult for them to use it with ease. Second of all, the instructions might have not been clear enough given that at least a couple of annotators did not use their intuition as native Spanish speakers, but rather their basic formal linguistic knowledge; therefore it cannot be said that these annotations were collected with a purely pragmatic approach but rather with a slightly mixed approach. Another reason why the instructions might not have been clear enough is that a couple of annotators reported the task to be quite difficult and at least one needed further clarifications. Given these important observations, the instructions for the final study were elaborated more thoroughly and more care was put into ensuring a correct experimental procedure.

An additional probable cause of the disagreement encountered might be the introduction of the label NaS. As shown in Section 3.5.1, even though this label was added as a mechanism to ensure the acceptability of the mood alternation pairs in the adverbal condition, given the high number of baseline pairs in which it was used and it's very sparse distribution (no pair was labeled as *not a sentence*), this label was clearly misused. Considering this, and the fact that the NaS label does not completely fit into the scale shown in Figure 3.1, in the final study this label was not brought into the analysis and was used as a filter.

Even though the aim should be to improve the experimental design and procedure to increase agreement among annotators, it should be noted that the *real* agreement is likely not too high due to several factors. First of all, the approach used here is pragmatic and, as already stated, a pragmatic approach means embracing uncertainty [3], which translates into lower agreement scores than in other settings. Second of all, as stated in Section 2.2.1, mood alternation is a phenomenon that reflects different presuppositions made about the information presented, and therefore, it is classified as a pragmatic phenomenon rather than a semantic one. Consequently, even more uncertainty and consequential greater disagreement should be expected from the annotations. Lastly, as the work from de Marneffe et al. [3] and that of Pavlick and Kwiatkowski [4] demonstrate, there appear to be cases in which the *true* label is split, that is, that to represent the judgments of speakers, not one, but two labels are required. In this study strong evidence that supports this idea has not been seen, but as stated in Section 3.5.5, more and better data might prove this.

Regarding the validity of the experimental design used for this pilot study, some changes can be made based on the effects seen and on the belief that a simplification of the design will clarify the results and therefore ease their understanding. The main change was to remove the specificity categories (common, mixed, and proper) given that no significant effect from them was detected in the model coefficients and no signs of possible effects were seen in the labels assigned. Another change was to remove the possibility condition. Even though it has proved to have a significant effect in the labels chosen, there are, to my knowledge, more resources and data for the negation condition, and also the tendencies seen in tables 3.10 and 3.11 are more compelling. Concerning the specificity conditions (individual and collective), even though there is no evidence to prove that they have a significant effect on the annotations, there are some signs of an influence that, with a better experimental procedure might become a significant or close to significant effect. Lastly, it was noticed in Section 3.5.5 that there was an important linguistic variable that was left out when defining the experimental design of this study, that of

the matrix verbs. Since this factor can influence the results, it must be included as a variable in the final study, but its exact definition will depend on the data gathered to build the corpus. With this we accomplish the second goal of this pilot study, ensuring the correctness of the experimental design and procedure.

As to the third goal, exploring different possibilities of statistical analysis, the main steps to be used in the final study can now be defined. First, the overall distribution of labels and the distribution of labels per batch will be examined, as well as calculate counts like those in Table 4.5, since all three have proven to be relevant to the posterior computations of inter-annotator agreement scores. As to these scores, the only one to be used in the final study is Gwet's AC_2 since it has proven to better reflect the annotations collected, and its implementation is settled. Lastly, given how informative the regression model has proven to be, a cumulative link mixed model will be fit to the data with probably a more thorough analysis of its characteristics. Aside from all this statistical analysis and as stated in the previous section, it should not be forgotten that what here was a very concise linguistic analysis, it needs to be more detailed for the final study.

Finally, concerning the first and second research questions, that is, whether the negation and possibility conditions affect the factuality of the embedded event when existing, the main event otherwise; based on the evidence presented here it can be said that they do affect the label assigned to the event, especially the possibility condition and with relative to the baseline, but not always significantly. Contrary to this, there is not enough evidence to state that our individual and collective conditions have a veridicality effect on the label assigned. But given the already mentioned problems in the experimental procedure, these answers should not be taken as final.

The next chapter presents the construction of the main or final study together with the analysis of the annotations gathered.

Chapter 4

Main Study

4.1 Introduction

In the previous chapter the pilot study, where a first answer to our research questions was given, was presented. The results showed an overall slight inter-annotator agreement ($AC_2 = 0.484$), which differed considerably when subsetting the dataset, since, for example, for the negation condition we had a slight agreement score of $AC_2 = 0.38794$. As to the distribution of the labels, a negatively skewed distribution was observed, with the positive labels being more used than the negative ones. Furthermore, an unexpected use of the label "not a sentence" (NaS) was seen. In addition, fitting a cumulative link mixed model (CLMM) yielded a significant difference between utterances with and without an adverb of doubt or possibility, with similar coefficients for the utterances with such an adverb where the verb is in the indicative mood and where the verb is in the subjunctive mood, and at least one more predictor with a coefficient close to being significant. Also from this model it was learned that the distance between the threshold coefficients, or, in other words, the space assigned for each label, was roughly equal, except for PR- and PS-. The last important result from the experiment is the observation of at least one more factor that affects the factuality judgments: the matrix verb.

Based on these results two important decisions were made for the main study presented here: to repurpose the NaS label and to simplify the experimental design. In the pilot study, the NaS label was used and analyzed as any other. Here, annotators could be used as any other, but instead of evaluating it with the others, it was decided to use it to filter annotations, as it will be clarified in Section 4.4. As to the simplification of the experimental design, the decision was made to facilitate the analysis of the results since for the pilot study it had proven cumbersome, even though it was not as deep as intended for this study. Specifically, it was decided to remove the specificity categories (common, mix, and proper) and the possibility or adverb condition. Consequently, the

research questions are now reduced to the following three:

RQ1.- In a complex sentence, how does the mood alternation of the embedded verb that occurs due to the negation of the main or matrix verb affect the factuality value of the embedded event?

RQ2.- How does an individual subject affect the factuality judgment of the event?

RQ3.- How does a subject that refers to a collective entity like an institution, affect the factuality judgment of the event?

From these questions we have the already known negation, individual and collective conditions; which can be grouped into specificity conditions (individual and collective) and mood alternation condition (negation). Furthermore, as in the pilot study the negation condition is divided into the categories of baseline, indicative, and subjunctive, exemplified in sentences 19a to 19c. The difference for the specificity conditions is indicated by /.

(19) a. El presidente/gobierno dijo que el
the.M.SG president.M.SG say.PST.PFV.IND.3SG that the.M.SG
país **tenía** problemas económicos.
country.M.SG **have.pst.ipfv.ind.3sg** problem.M.PL economic.M.PL
"The president/government said that the country **had** economic problems."

b. El presidente/gobierno no dijo que el
the.M.SG president.M.SG not say.PST.PFV.IND.3SG that the.M.SG
país **tenía** problemas económicos.
country.M.SG **have.pst.ipfv.ind.3sg** problem.M.PL economic.M.PL
"The president/government didn't say that the country **had** economic problems."

c. El presidente/gobierno no dijo que
the.M.SG president/government.M.SG not say.PST.PFV.IND.3SG that
el país **tuviera** problemas
the.M.SG country.M.SG **have.pst.ipfv.sbjv.3sg** problem.M.PL
económicos.
economic.M.PL
"The president/government didn't say that the country had economic problems."

To obtain the experimental design, these conditions are crossed as in the pilot study, obtaining, in this case, a 3×2 design shown in Table 4.1, where we also see that the terminology used before of seeds and variants is kept here. Although in Section 3.5.5 it

Experimental Conditions			
		SPECIFICITY	
MOOD-Negation		Individual	Collective
	Baseline	S1	S2
	Indicative	V1	V2
	Subjunctive	V1	V2

TABLE 4.1: Experimental conditions for the main study. As in the pilot study, **S** followed by a number stands for a set of seeds, **V** stands for variants, with the number indicating the corresponding set of seeds.

was proved that the veridicality of the matrix verb affects the factuality of the embedded predicate and thus the label chosen by the annotator, it is not included here directly in the experimental design but rather as information annotated, as it will be explained in the next section, and tracked it closely in the analysis of the annotations. This was done to prevent a more complicated experimental design.

4.2 Dataset

Different sources were used to create the corpus. First, to directly compare results with the pilot study, we used all pairs from the negation condition. Then, candidate premises were extracted from a section of the Daves Corpus del Español [43], the Old News Corpus for Spanish [44], *El Quijote* by Miguel de Cervantes (as found in Dario [45]), the XNLI corpus [16], and the United Nations corpus in Spanish for the years 2000, 2001, 2002, and 2003 [46]. The candidate premises were sentences in the indicative condition whose embedded verbs were not in the simple future or conditional tense and with no specific matrix verbs as in the pilot study. These candidate sentences were found thanks to the LinguaKit [47]. After that, the following modifications were done:

- Premises were shortened to be under 30 tokens, when necessary.
- Additional verbs in personal forms were removed.
- References were resolved when it was both needed and clear. There was at least one case in which it was needed, but it wasn't clear, and therefore an arbitrary substitution was made.
- Collocations like "to be a doubt" were not included.

- In a few cases where the verb was in first person (singular or plural) the entity to which the subject refers was just put ahead of the premise in the following format: *Lotario: I think that the sky is blue."*
- In one case the aspect of the embedded verb was changed to allow mood alternation since there are no simple forms in perfect aspect for the subjunctive.
- Adverbs modifying the matrix verb were removed.
- In a few cases, the subject was modified so the premise could be used both for individual and collective conditions.

Once these modifications were done, the hypotheses were extracted and the pairs were modified following the experimental design. Then different lexical, morphological, and statistical labels for each pair were gathered with the idea of maybe using them to extend the analysis of the crowdsourced annotations. The most relevant features gathered are the lexical items corresponding to the matrix and the embedded verbs, the veridicality value of the matrix verb, and person, number, tense, and mood for both the matrix and the embedded verb. So for example (19a in either of the specificity conditions part of the information we have is the following: { matrix: decir, auxiliary: NaN, modal: NaN, veridicality: o/o[1], length_premise: 9}

It should be noted though that the distribution for most of these labels is quite sparse and that the veridicality annotations, although done with the help of Stanford [48], are my own annotations and they do have a significant degree of uncertainty. Lastly, to obtain an overall view of the corpus, counts of these features and some basic statistics were computed. Table 4.2 shows a sample of these.

4.3 Predictions

Given that several of the results obtained in the pilot study were unexpected and not completely explained, it was not considered important to define new detailed predictions and thus those from the pilot study, repeated with the needed adaptations in Table 4.3, are kept as guidelines for this experiment, although in this case deviations from them are certainly expected since, as mentioned in the previous chapter we know, for example, that verb matrices affect the factuality of the hypotheses.

[1]Neutral in affirmative and negative contexts.

Basic Corpus Statistics	
	Value
#pairs	524
Average premise length	15.004
Average hypothesis Length	9.539
Most Frequent Matrices	
saber (to know)	102
creer (to believe)	48
considerar (to considerate)	45
decir (to say)	36
anunciar (to announce)	36

TABLE 4.2: Some basic corpus statistics.

Label Predictions within Experimental Design				
			SPECIFICITY	
MOOD	Negation		Individual	Collective
		B	CT+	CT+
		I	PR+	PR+
		S	PS-	PS-

TABLE 4.3: Rough approximation of label predictions for each combination of experimental conditions. **B, I, S** stands for *baseline, indicative, subjunctive*. The only difference from the pilot study is the removal of the no longer used possibility condition and the specificity categories.

Aside from these label predictions, it is important to remark that, while creating the corpus it was noticed that some pairs could be *problematic*, that is, pairs for which disagreement or unexpected labels was considered possible. A deeper explanation of these factors will be given in Section 4.6.5, so for the moment just mention that some of the factors considered are the nature of the hypothesis, like *Las mujeres son seres humanos* (women are human beings), indefinite determiners like *algunos* (some), and the type of utterance that the premise is, like assertions vs. questions.

4.4 Procedure

Annotations were collected on the platform Toloka [13] with the same labels as in the pilot study (see Figure 3.1). The interface used was similar to that of the pilot study with the difference being that the labels were ordered from positive (CT+) to not a sentence (NaS). Figures 4.1 and 4.2 show what it looked like. Workers were paid $0.433 per set of pairs, they were required to be Spanish speakers and were chosen among countries where Spanish is either an official language or one of the most important unofficial ones. Additionally, after the first sets of annotations were gathered it was decided to choose

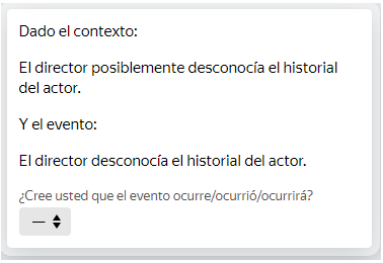

FIGURE 4.1: Interface for annotations

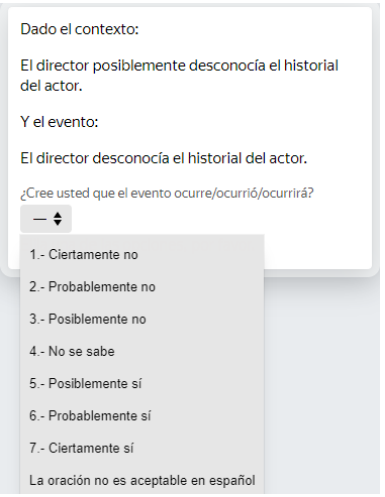

FIGURE 4.2: Interface with labels

only the top 50% annotators, and for the final sets, the top 30% annotators. No batches were used since there was no way to ensure annotators did not cross between them, but pairs were gathered in sets of no more than 10, and annotators were allowed to skip them. As to the number of annotators per pair, as many as the budget allowed were gathered.

While running the experiment, annotations were reviewed by sets of submitted pairs. If the set was submitted too fast, like less than 50 seconds for 10 pairs, or if most of the pairs had the same label (9 out of 10 pairs, for example), it was usually rejected. Train or control tasks were not added since it was considered that they could lead annotators towards specific labels, or in other words, towards a lexical approach rather than a pragmatic one.

Once all annotations were collected, following Pavlick and Kwiatkowski [4] all pairs which were labeled as *not a sentence* by at least one annotator were dropped, and if a worker annotated more than 1 pair within a single combination of experimental conditions, all his annotations in that combination were removed. This left a total of 477 pairs and 7 annotators per pair. These are the annotations whose analysis is presented here.

Use of NaS	
	Count
Total	66
# pairs	63
# pairs # NaS > 1	3
# NaS baseline	20
# NaS indicative	19
# NaS subjunctive	24
# NaS individual	36
# NaS collective	27

TABLE 4.4: Basic counts on the use of the label *not a sentence* (NaS).

4.5 Analysis of NaS

Before presenting the results of the corpus, a brief analysis of the use of the label NaS is presented. Although there is previous work on discarding sentences that one annotator considers as not acceptable, given its unexpected use in the pilot study and the decision to change its purpose, seeing if there are any underlying patterns that could explain this rejection, is worth the effort. Table 4.4 shows some counts of this label.

Even though for this experiment there was no explicit construction whose acceptability was doubted and the use of this label was explicitly discouraged in the instructions, it was still used in a considerable amount of cases, which suggests that it was a conscious decision. But, as in the pilot study, its use is quite spread across the pairs, which questions the exact reason why annotators chose it.

As to the distribution among experimental conditions, it was expected that the use of NaS would have similar behavior as to the disagreement found in the pilot study because it was assumed that non-acceptability is one of the causes of disagreement. But as we can see on the table, this does not seem to be the case. For example, in the pilot study , there was more disagreement in the collective condition than in the individual one, but here NaS was used more frequently in the individual condition. More importantly, NaS was chosen quite frequently in the baseline condition, which should not be problematic.

In conclusion, not forgetting that the amount of data analyzed here is small and that the analysis itself is shallow, there seems to be no pattern that explains the use of the label *not a sentence* (NaS), thus it appears that there is no support to having this label or to using it as a filter. Next, the analysis of the corpus is presented.

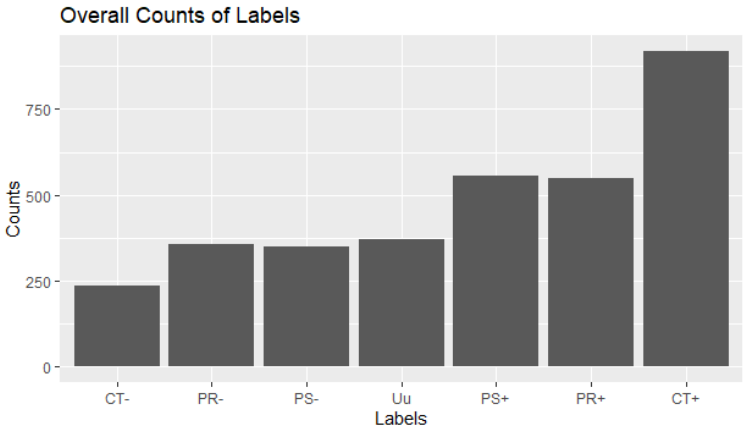

FIGURE 4.3: Overall distribution of the proportion of labels used by annotators.

4.6 Results

4.6.1 Overview of the Annotations

Figure 4.3 shows the distribution of label counts for this experiment. As for the pilot study, the distribution is negatively skewed, but there are also some important differences. First of all, the frequencies for the negative labels are relatively higher than for the previous case. This is probably due to the removal of the possibility condition and the more diversified matrices. Another interesting distinction relating to the negative labels is the use of the label *certainly not*. In the pilot study, negative labels were roughly equally used, at most, the use for CT- was slightly higher, but in this case, the use for this label is lower than the other negative labels. But more important is the situation of the *unknown or uncommited* label, which, contrary to the previous case, is considerably lower than the positive labels. Lastly, it should be noted that here PR+ and PS+ are almost equal, suggesting that the review of one of the annotators of the pilot study (that he couldn't distinguish between them) is now confirmed.

To further understand the overall features of the annotations some very basic computations, shown in Table 4.5, were done. Regarding the agreement patterns summarized by the counts of pairs with > 3 votes on one label or majority label, with a unique label but < 3 votes or most votes label, and without a unique label; they indicate that the inter-annotator agreement score is likely to be quite low, but since for 80% of the pairs there is 1 label, there might some underlying tendencies that the experiment has not flashed out. Next, we will go over the inter-annotation agreement scores.

Basic Statistics		
	Count	Percentage
Number of analyzed pairs	477	88.333
Number of dropped pairs	63	11.6
Number of analyzed annotations	3339	-
Total number of annotators	248	100
Average annotations per worker	13.464	-
Standard deviation annotations per worker	15.261	-
Maximum #annotations per worker	98	-
Minimum #annotations per worker	1	-
#pairs with > 3 votes on one label	195	41.139
#pairs with a unique label but < 3 votes	201	42.405
#pairs without a unique label	78	16.456

TABLE 4.5: Some basic counts of the annotations.

4.6.2 Inter-Annotator Agreement Scores

Overall the inter-annotator agreement score for the whole corpus is $AC_2 = 0.114$ which is barely within the range of slight agreement $(0.11 - 0.40)$ [31] and which is far away from the pilot's study $AC_2 = 0.388$ for the negation condition. Given this, it was decided to explore the value of Gwet's Ac_2 for different subsets of the corpus to try to find what causes such big disagreement. Subsetting by veridicality values or by specificity conditions did not yield very informative results, but, as seen in Table 4.6, subsetting by mood categories or by whether the pair was or not in the pilot study gave some interesting results.

The biggest difference seen is between the baseline and the mood alternation conditions, there is a drop from fair agreement to virtually none $(0.00 - 0.10)$, which can be considered as evidence that the baseline is well established. Nevertheless although considerably higher than for the whole corpus, the agreement for the baseline category is still lower than for the pilot study, which at least partially explains why for the other two conditions the score is quite low since mood alternation is a pragmatic phenomenon and thus more uncertainty with relative to the baseline is expected. Furthermore, the difference between the indicative and the subjunctive is minimal, which can be regarded as a sign that mood alternation does not affect the factuality of the embedded event, or at least barely so.

A very puzzling result shown in the table is the difference between the pairs that were in the pilot study and those that were not. Since an important difference between the pairs in both studies is that the matrix verbs in the pilot are the standard matrices for

Inter-Annotator Agreement Score for Different Subsets	
Subset	AC_2
ALL	0.114
Baseline	0.194
Indicative	0.070
Subjunctive	0.085
Pairs in pilot study	0.080
Pairs not in pilot study	0.129

TABLE 4.6: Inter-annotator-agreements scores for the whole corpus and different subsets.

Inter-Annotator Agreement Score for Different Matrices	
Subset	AC_2
ALL	0.114
Top 5 most frequent matrices	0.123
Not top 5 matrices	0.093
saber (to know)	0.170
creer (to believe)	0.131
considerar (to considerate)	0.098
olvidar (to forget)	0.181
ver (to see)	0.022

TABLE 4.7: Inter-annotator-agreements scores for the whole corpus and different subsets by matrix verb. The 5 most frequent matrices are *saber* (to know), *creer* (to believe), *considerar* (to considerate), *decir* (to say), and *anunciar* (to announce).

mood alternation according to Real Academia Española [23] and those in this study are *natural occurrences*, the corpus was subsetted according to whether the matrix verb is standard or not, but the difference was minimal. Thus the difference between being or not in the pilot study probably does not have a straightforward explanation.

Aside from the already mentioned subsets, AC_2 was computed on different matrix verbs subsets and the results are presented in Table 4.7. At first, it was suspected that rare matrices could cause disagreement and therefore the corpus was divided into two groups: pairs whose matrix verb is one of the 5 most frequent ones (*saber* (to know), *creer* (to believe), *considerar* (to considerate), *decir* (to say), and *anunciar* (to announce)) and pairs which do not. This yielded a small difference in the score that was explored even further with specific matrices with different frequencies. By doing so, it was proved that the variations seen with different matrices are not due to the frequency of the matrix but rather to the matrix itself. For example, *saber* (to know) has a frequency of 102 and an AC_2 score of 0.170, but, *olvidar* (to forget), whose frequency is 24, lower than *saber* (to know) and even than *creer* (to believe, 48), has an AC_2 of 0.181. In conclusion, it seems that the lexical nature of the matrices influence agreement. Next, the results of

Model Scores							
link	Threshold	Nobs	logLik	AIC	niter	max.grad	cond.H
logit	flexible	3339	−6179.67	12379.34	964(1997)	7.36×10^{-03}	1.4×10^{02}

TABLE 4.8: Cumulative link mixed model scores for the model representing the whole dataset.

fitting a cumulative link mixed model (CLMM) are presented.

4.6.3 Model Fitting

Several CLMMs with the label as an outcome variable were fitted for both the whole dataset and the datapoints corresponding to the indicative and subjunctive categories. Annotator and premise-hypothesis pair were set as random effects, with the former barely changing within the different models and the latter showing much greater model variability. As predictors the mood condition, the specificity condition, the different matrix values, and the different veridicality values were used. Predictors were either single or crossed. The maximum number of crossed predictors was 4. AIC values differ, but not greatly: for the whole dataset values were all between 12300 and 12500 , with differences smaller than 120; for the indicative-subjunctive dataset values were between 8300 and 17500. Conditional Hessian values range from 10^2 to 10^5.

From these different models, there are some important observations to be made. First of all, the specificity conditions complement other predictors in a few cases, that is, they grade their information rather than being informative on their own. Secondly, mood appears to be more relevant than specificity since it can yield significant effects, but when combined with other predictors it grades them or completely disappears. More importantly, for the subset corresponding to the indicative and subjunctive categories, there is only one case in which mood has a significant effect: for the matrix verb *olvidar* (to forget). That is, there is a difference between the baseline and the mood alternation conditions, but not between the latter, suggesting that the mood alternation phenomenon in its negation instance does not affect the factuality of the embedded verb significantly. Lastly, given the high conditional Hessian values and the low pair variance reached with the predictors already mentioned, adding more from the information gathered from each pair would be a mistake.

Model Coefficients			
Coefficient	Estimate	Standard Error	$\Pr(>\|z\|)$
MOOD_CONDITION: indicative	-0.372	0.083	8.32×10^{-06}
MOOD_CONDITION: subjunctive	-0.438	0.084	1.71×10^{-07}

TABLE 4.9: Cumulative link mixed model coefficients for the model representing the whole dataset.

Model Random Effects			
Groups	Name	Variance	Std.Dev.
PAIR	(Intercept)	0.101	0.318
RATER	(Intercept)	0.010	0.101

TABLE 4.10: Cumulative link mixed model random effects for the model representing the whole dataset.

To represent the whole dataset, a model with the formula LABEL = MOOD_CONDITION + (1 | ANNOTATOR) + (1 | ID) was chosen, and its main results are presented in tables 4.8 to 4.11. This model was selected based on the scores presented in Table 4.8 and given that the mood conditions are one of the main variables of this study. Adding the specificity conditions proved to be non-informative and adding the matrices increased both AIC and the conditional Hessian matrix considerably, which is suspected to be caused by their sparse distribution. A similar effect to a lesser degree occurred with the conditional Hessian matrix when adding the veridicality values, and since these values are also quite sparsely distributed, this model was also dismissed.

Table 4.9 presents the model coefficients. Both indicative and subjunctive are significantly different from the baseline, thus showing that overall the baseline is well established. Furthermore, both coefficients are negative and rather small, proving the notion expressed in the experiment's predictions in Table 4.3: the factuality of the hypothesis decreases with the negation of the matrix and that overall there is a difference between having the embedded verb in the indicative or in the subjunctive mood. Concerning this difference, it is not so surprising that the models fitted only to the indicative and subjunctive conditions proved insignificant given that the difference between the coefficients in Table 4.9 is < 0.1.

From Table A.1.2, which presents the random coefficients of the model, the most interesting result is that both random effects are considerably lower than for the pilot study (0.512 and 0.505 respectively). Given the ample disagreement found in this study, one could have expected it to be reflected in these random variables. In other words, if there was a lot of noise in the data due to a poor collection in the annotations, this would have meant a lot of variance in the annotator variable. Since this is not true, it

Model Threshold Coefficients			
Threshold	Estimate	Std. Error	z value
CT-—PR-	−2.910	0.089	−32.564
PR-—PS-	−1.851	0.072	−25.534
PS-—Uu	−1.239	0.067	−18.453
Uu—PS+	−0.723	0.064	−11.238
PS+—PR+	−0.027	0.063	−0.425
PR+—CT+	0.724	0.064	11.301

TABLE 4.11: Cumulative link mixed model threshold coefficients for the model representing the whole dataset.

seems that the experimental procedure is acceptable and that what should be improved is adding more predictors that could reduce the pair variance. A more balanced corpus could solve this.

As to the model threshold coefficients presented in Table 4.11, we can see that the spaces estimated for each label roughly range within $(0.5 - 1.2)$, which means that they are smaller than in the pilot study, where distances had a rough range of $(0.3 - 1.5)$. Furthermore, the differences between negative labels and positive labels have been reduced concerning the previous experiment. These differences concerning the pilot study are broadly consistent with the differences seen in the labels' distribution.

Lastly, to understand whether the sparsed distribution of the matrices explains why they do not work well as a predictor, the same model and a model that adds matrix as a predictor were fitted to the subset with the five more frequent matrices where matrix frequencies range from 36 to 102 and their results are presented in the appendix. AIC values were almost identical and the model with fewer predictors had a smaller value for the conditional Hessian matrix. Nevertheless, for the second model, the conditional Hessian value was reduced by more than a 100, and more significant effects ($p < 0.05$) were yielded than the model with the same formula for the whole dataset. This confirms what was mentioned above: with a more balanced corpus more predictors that explain the pair variance can be added.

Next, an analysis of the different agreement patterns is presented. In other words, an analysis of how votes in each pair are distributed is presented.

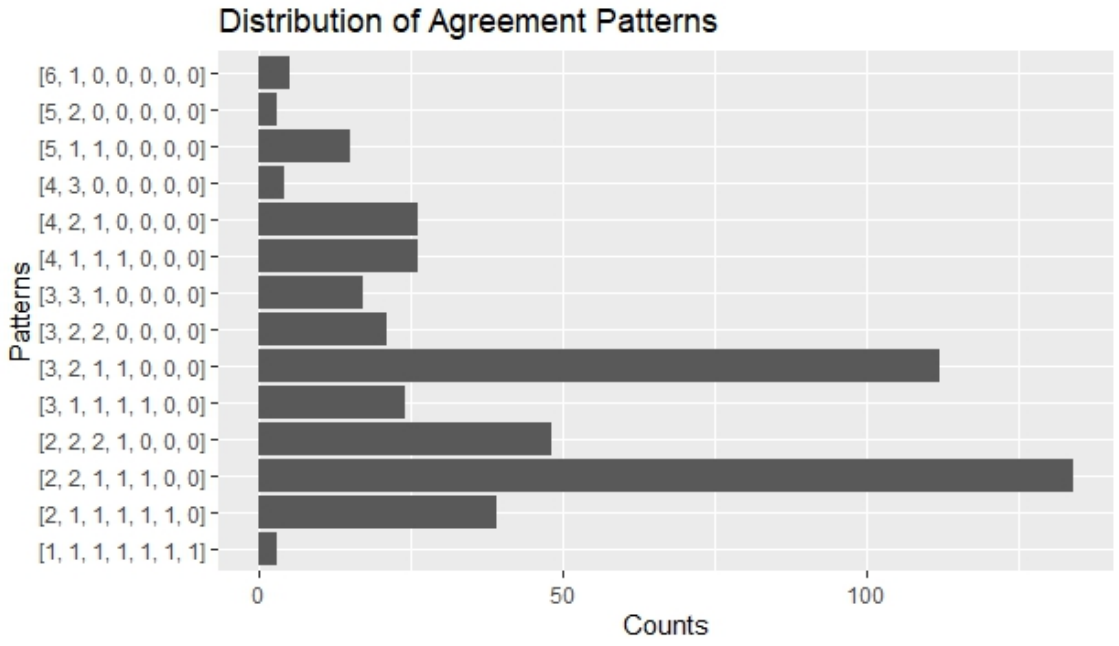

FIGURE 4.4: Distribution of agreement patterns.

4.6.4 Agreement Patterns

Given the amount of disagreement found in this study, it seems futile to try to do as in the pilot study and generate a table with the assigned label/s per each combination of experimental combinations. Thus it was decided to instead explore the agreement patterns and see what information they can yield.

Figure 4.4 shows the distribution of the different agreement patterns found in the corpus. Although there is a considerable amount that occurs less than 25 times ($< 6\%$), there are some patterns that occur with a distinguishable frequency which show the lack of agreement measured by the overall AC_2 score. Among these, there are two patterns that more than double the frequencies of the other agreement patterns and thus cannot be disregarded: $[3, 2, 1, 1]$ and $[2, 2, 1, 1, 1]$. Although neither of them can be mapped to an exact label split, that is, it cannot be said that, for example, they correspond to having 2 or 3 labels; it is certainly evidence of the existence of inherent disagreement, as in the study of Pavlick and Kwiatkowski [4].

If we look at how the agreement patterns are distributed per mood condition (Figure 4.5), we can see that there are some differences, although the picture is not quite clear, and that the results are consistent with the agreement scores in Table 4.6 since within

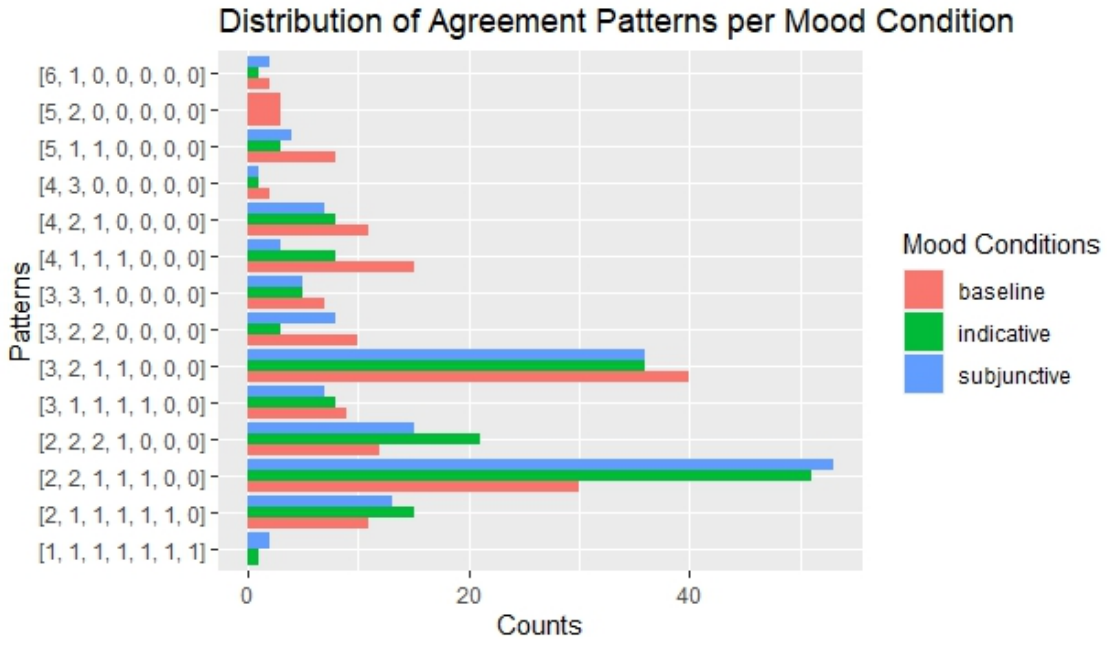

FIGURE 4.5: Distribution of agreement patterns per mood condition.

each pattern, pairs are more often in the baseline condition for the patterns that represent greater agreement and pairs are in the indicative and the subjunctive more often for the more scattered patterns. Specifically, the baseline condition has the highest frequency from $[6, 1]$ to $[3, 1, 1, 1, 1]$, and the indicative and the subjunctive are in close competition in patterns from $[2, 2, 2, 1]$ to $[1, 1, 1, 1, 1, 1, 1]$. This suggests that the pairs in the indicative and subjunctive might be better represented with 3 labels, but for the pairs in the baseline condition frequency, this suggests that some cases could be better represented with 2 labels.

It is quite surprising that there is such a considerable amount of disagreement for pairs that do not have the main linguistic markers in this study, neither negation adverb nor mood alternation. It is true that given this lack of agreement is not surprising that the overall agreement is barely within the range of fair agreement. Given this, it is worthwhile to look closer at this condition to try to understand what happened. To do so, we look at all baseline pairs which have *negar* as a matrix verb, 6 in total. Out of these, there is majority agreement for only one pair, three have an agreement of the type most votes, and two do not have an agreement upon a unique label. We focus on these last two pairs.

(20)	a.	**Su**	Gobierno	niega	que	la
		His/her/its.sg	government.M.SG	deny.PRS.IND.3SG	that	the.F.SG
		aplicación	práctica	de **esas**	disposiciones	
		application.F.SG	practical.F.SG	of **that.f.pl**	disposition.F.PL	

deja mucho que desear.
leave.prs.ind.3sg much that desire.inf
His/her Government denies that the practical application of **those** dispositions **leaves much to be desired**.

b. La aplicación práctica de
the.F.SG application.F.SG practical.F.SG of
esas
that.f.pl disposition.f.pl leave.prs.ind.3sg much that desire.inf
disposiciones **deja mucho que desear**.

The practical application of **those** dispositions **leaves much to be desired**.

Pair (20) seems to be in the realm of uncertainty (PS+ to PS-) since the two most voted labels are PS+ and PS- with two votes each, and there is only one vote for a certain label, specifically CT+. Having negative labels (3 in total) is expected since in positive environments *negar* (to deny) makes the embedded predicated a contradiction. As to why the judgments lean towards uncertainty more than negative certainty two factors might have some influence: reference solution and the relativity associated with the embedded predicate. Regarding the first factor we have two unresolved references, *su* (his/her) and *esas* (those) that could have caused the workers to wonder whose government and which dispositions is the premise referring to, and thus made the judgment about the factuality of (20b) more uncertain. As to the second factor, *dejar mucho que desear* (to leave much to be desired) is an at least partially subjective expression since what one desires often depends from person to person. Thus when assessing the truthfulness of the hypothesis is likely that the annotators added another layer of uncertainty but, contrary to the previous one, I believe this one depends more on the annotator, and thus more than explaining the uncertainty, it explains the lack of agreement, whereas the layer created by the reference markers could be the cause the most voted labels being PS+ and PS-. But of course, a more thorough analysis is needed to validate this hypothesis.

(21) a. Israel niega que en el ejercicio de su
Israel deny.PRS.IND.3SG that in the.M.SG exercise.M.SG of his/her/its.SG
derecho inherente a defenderse del terrorismo
right.M.SG inherent.SG to defend.INF.REFL from.the.M.SG terrorism.M.SG
más brutal **debe** actuar dentro de los límites
most brutal.SG **must.prs.ind.3sg** act.INF inside of the.M.PL limit.M.PL
del derecho internacional.
of.the.M.PL law.M.SG international.M.SG

Israel denies that in the exercise of its inherent right to defend itself from the most brutal terrorism **must** act inside the limits of international law.

b. En el ejercicio de su derecho inherente a
in the.M.SG exercise.M.SG of his.her.its.SG right.M.SG inherent.SG to
defenderse del terrorismo más brutal Israel
defend.INF.REFL from.the.M.SG terrorism.M.SG most brutal.SG Israel
debe actuar dentro de los límites del
must.prs.ind.3sg act.INF inside of the.M.PL limit.M.PL of.the.M.PL
derecho internacional.
law.M.SG international.M.SG
In the exercise of its inherent right to defend itself from the most brutal terrorism Israel **must** act inside the limits of international law.

As to pair (21) the situation is quite different. To begin with, we have a 3-way split between PS+, PR+, and CT+; which, considering the already mentioned behavior of *negar* (to deny), is quite surprising. In this case, we do not have any unresolved references or subjective predicates, but there is an embedded deontic modal predicate, *debe actuar* (must act), which makes the entire hypothesis an obligation [7], as it was seen with *tener que* (must) in Chapter 1, and thus leaving its acceptance or rejection in the world of possibilities and consequently less certain. In other words, the presence of the deontic modal could be the cause of the label split which is consistent with evidence shown by Benamara et al. [49] on the impact of deontic modals on sentiment analysis in French. As to the reason behind most of the labels being in the positive range, it is likely to be the world knowledge associated with the utterance since the rule of international law is usually accepted, even when having to defend yourself. Consequently overriding the effect of the matrix verb. More examples of how world knowledge affects judgments are presented in the next section.

The analysis presented for pairs (20) and (21) is not enough to explain the amount of disagreement found in the baseline category, but it provides some evidence that other factors complicate what was supposed to be a *simple* assessment. A more thorough and formal analysis is left for future work.

Lastly, within this section, we have Figure 4.6, which presents the distribution of agreement patterns according to the two specificity conditions: individual and collective. The overall picture appears to be quite confusing with not a very clear sign of for which condition there is more agreement. By distributing the agreement patterns according to their agreement type (majority, most votes, not unique), we see that there are more pairs in the collective condition that have majority agreement, that is, one label with

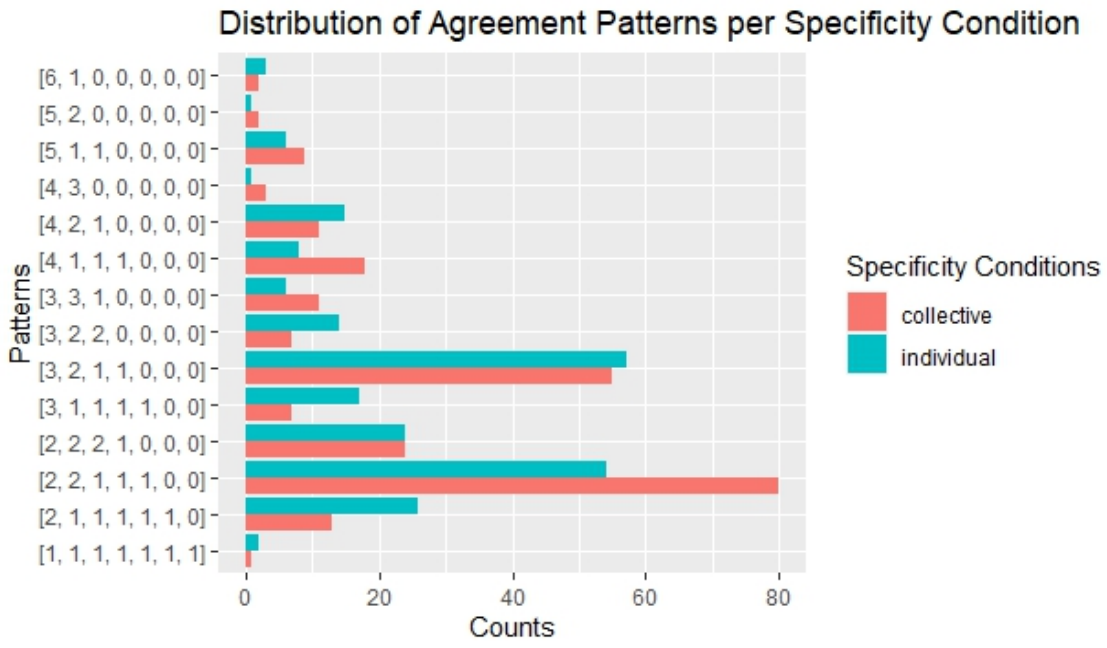

FIGURE 4.6: Distribution of agreement patterns per mood condition.

at least 4 votes. But, given that patterns of type most votes (one label has most of the votes but less than 4) are more often in the individual condition, and that in the pattern with the highest frequency, there are more pairs in the collective condition, we cannot confirm the observation made in the pilot study that disagreement is higher for the collective condition. Figure 4.6 rather shows that there are no big differences between the specificity conditions, which is consistent with the agreement and the model's results.

4.6.5 Manual Analysis

The goal of this section is to take a closer look at the annotations gathered to search for more causes of disagreement and to see if the *problematic* pairs mentioned in Section 4.3 are indeed problematic or not. Before analyzing the pairs it should be noted that to avoid long lists of glossed examples only pairs in the baseline conditions, that is individual or collective seeds, are represented here; given the results on the specificity conditions not much attention will be paid to them, and more importantly, that this analysis is only exploratory. A systematic analysis is left for future work.

In the entire corpus, no seed was found that *literally* fits the predictions made in Table 4.3, but some examples show small variations between the different mood conditions and a few are close to the predictions. The best example is the pairs represented by seed

(22), which were labeled as CT+, PR+, and PS+, with 4, 3, and 5 votes respectively. The matrix verbs for these pairs, *anunciar* (to announce), is neutral in both affirmative and negative contexts, that is, the premise neither entails nor contradicts the embedded predicate no matter the polarity of the whole sentence. This, together with the absence of any other linguistic markers that could interfere with the mood conditions and that it is unlikely that the hypothesis generates strong presuppositions or implicatures, explains the lack of strong deviations from the predictions.

(22) a. La familia anunció que
the.F.SG family.F.SG announce.PST.PFV.IND.3SG that
había recortado gastos.
have.PST.IPFV.IND.3SG cut.down.PTCP.M.SGexpense.M.PL
The family announced that they had cut down expenses.

b. La familia había
the.F.SG family.F.SG have.PST.IPFV.IND.3SG
recortado gastos.
cut.down.PTCP.M.SGexpense.M.PL
The family had cut down expenses.

The label assigned for the subjunctive condition is interesting because it suggests that the effect of the negation adverb is not very strong, less than expected. Although a more thorough analysis to fully understand the impact of *no* (not) in this context.

A very interesting example is (23), for which the label CT+ was assigned in each of the conditions with more than 3 votes, more specifically with 4, 6, and 6 votes respectively. That is, on average the factuality of the event *las mujeres son seres humanos* (women are human beings) is not modified either by the negation of the matrix verb or by the mood alternation of the event itself. This could indeed be partially explained by the fact that *saber* (to know) implies the embedded predicate in both affirmative and negative contexts, but given that not all pairs that have this matrix verb have these labels, it is likely that the deciding factor is the knowledge associated to the hypothesis. The event represented in the hypothesis is a state whose negation or even just uncertainty is not likely to be accepted in many societies nowadays. In other words, it is likely that the world **knowledge of** [many of] the speakers overrode all linguistic factors at play.

(23) a. Algunos delegados gubernamentales saben que
some.M.PL representative.M.PL governmental.M.PL know.PRS.IND.3SG that
también las mujeres son seres humanos.
also the.F.PL woman.F.PL be.PRS.IND.3PL being.M.PL human.M.PL

Some governmental representatives know that women are also human beings.

b. Las mujeres son seres humanos.
the.F.PL woman.F.PL be.PRS.IND.3PL being.M.PL human.M.PL
Women are human beings.

Another good example of how world knowledge affects factuality judgments is pair (24). In the baseline and indicative conditions this seed received a majority vote for CT+ with 4 votes in each case, and votes for the subjunctive condition were split between PR- and CT+ with 2 votes each. Aside from being in the collective condition and the respective mood conditions, we have on one hand *negar* (deny), which contradicts the embedded predicate if it is not negated, as a matrix verb, and on the other hand we have *poder* (can), which is here functioning as a **deontic modal** and therefore, as the other deontic modals already presented even if of a different kind, puts both *negar* and the embedded predicate in the realm of possibilities. Nevertheless, it seems that for most of the annotators, neither of these factors was significant, but instead their world knowledge overrode them since *los jóvenes tienen relaciones sexuales* (young adults have sexual relationships) is often considered an undeniable fact. However, it appears this *universal truth* is not shared by most of the annotators since the same number of annotators that consider the hypothesis as an undeniable fact in the subjunctive condition labeled it as *probably not*. That is, world knowledge can override linguistic features but it is difficult to determine when it is going to happen.

(24) a. Los gobiernos **pueden** negar que los
the.M.PL government.M.PL **can.prs.ind.3pl** deny.INF that the.M.PL
jóvenes tienen relaciones sexuales.
young.adult.M.PL have.PRS.IND.3PL relationship.F.PL sexual.PL
The governments **can** deny that young adults have sexual relationships.

b. Los jóvenes tienen relaciones sexuales.
the.M.PL young.adult.M.PL have.PRS.IND.3PL relationship.F.PL sexual.PL
Young adults have sexual relationships.

In Section 4.3 it was mentioned that some of the pairs in the corpus were questions. Specifically, two seeds were added in the individual condition with certain concerns. This concern came because it was assumed that questions would add another level of uncertainty to an already uncertain ground. Thus we will now see what the annotations for these pairs are.

The seeds for these **questions** are presented in examples (25) and (26). Both pairs are quite similar since they have the same matrix verb *saber* (to know) in the same tense

and aspect, the embedded events are states of different nature but in the same tense, and they even have the same **adverb** *acaso* (by any chance), which should have been removed when creating the corpus, modifying the matrix verb. As to the labels obtained, they are quite interesting since, except for (25) in the subjunctive condition, neither of the variants agree upon one label, although the number of splits and the labels in which they split is different.

(25) a. **¿Acaso** sabe el Sr. Bustani que Israel
by.any.chance know.PRS.IND.3SG the.M.SG Mr. Bustani that Israel
está armado hasta los dientes con armas
be.PRS.IND.3SG arm.PTCP to the.M.PL tooth.M.PL with weapon.F.SG
nucleares y con las aeronaves y los tanques más
nuclear.PL and with the.F.PL aircraft.F.PL and the.M.PL tank.M.PL more
avanzados?
advance.PTCP.M.PL
Does Mr. Bustani know, **by any chance**, that Israel is armed to the teeth with nuclear weapons and with the more advanced aircrafts and tanks?

b. Israel está armado hasta los dientes con
Israel be.PRS.IND.3SG arm.PTCP to the.M.PL tooth.M.PL with
armas nucleares y con las aeronaves y los
weapon.F.SG nuclear.PL and with the.F.PL aircraft.F.PL and the.M.PL
tanques más avanzados.
tank.M.PL more advance.PTCP.M.PL
Israel is armed to the teeth with nuclear weapons and with the more advanced aircrafts and tanks.

In the baseline condition (25) was split into PS+/PR+ and (26) into PR-/Uu/CT+, in both cases with 2 votes per label. Given the already mentioned confusion between probability and possibility, the split in the first case is not surprising; and if to that we add the uncertainty assumed for these questions and the effect of the adverb, the chosen labels are both acceptable and explainable. As to the second case is unexpected, but when looking directly at the hypothesis (26b), it seems that it might be the cause of the disagreement. *Cada persona es un microcosmos* (each person is a microcosmos) is a statement of a philosophical nature, thus some people consider it an *undenniable* truth while others might negate its factuality. Either way, this kind of predicates likely generates strong presuppositions that are more likely to override linguistic markers than other types.

(26) a. ¿**Acaso** sabemos todos que cada persona
by.any.chance know.F.SG everyone that each.F.SG personF.SG
es un microcosmos?
be.PRS.IND.3SG a.M.SG microcosmos.M.SG
Do we all know, **by any chance**, that each person is a microcosmos?

b. Cada persona es un microcosmos
each.F.SG personF.SG be.PRS.IND.3SG a.M.SG microcosmos.M.SG
Each person is a microcosmos.

As to the results for the other two conditions of the seed (26), the explanation holds since the range of labels used is the same, although the splits change. Regarding seed (25), there is more variation. In the indicative condition, we have again a 2-2 split, but instead, we have the labels Uu and CT+. For the subjunctive condition, there are no splits at all, instead, we have a majority agreement for CT+. The only reasonable explanation is that there is a *struggle* between the linguistic markers and workers' knowledge about the truthfulness of *Israel está armado hasta los dientes con armas nucleares y con las aeronaves y los tanques más avanzados* (Israel is armed to the teeth with nuclear weapons and with the more advanced aircrafts and tanks), and in the subjunctive condition annotators favored the factuality of this hypothesis, which is considered an undeniable truth by some, over the linguistic markers, similar to the results for (24).

Regarding the pairs with indefinite determiners, specifically with *algunos* (some), there were no clear patterns found overall. If they cause uncertainty, it seems that it is to a lesser degree in comparison with questions, since only 2 pairs out of 12 had split labels. Only 3 pairs indeed reached a majority agreement, but the labels do not reflect a clear lack of confidence in the judgments, especially with the results already explained for seed (23).

Lastly, after the results of pair 21, it was hypothesized that as we have seen with modals modifying matrix verbs, the presence of modals in embedded predicates could also be a cause for uncertainty or disagreement. Therefore the pairs where there is a modal verb in the embedded predicate were examined, a total of 30 pairs, and it was found that for half of them, there is no unique label upon which annotators agreed, a percentage quite higher than the overall 16.456%. Furthermore, 4 out of these 15 pairs were split into three labels. Thus, the presence of a deontic modal like *deber* (must) causes disagreement. But of course more data is needed to confirm this hypothesis.

To sum up, in this section we have presented signs that explain the disagreement and the unpredicted labels found for some pairs. Of these, given their reiterative appearance, the most important is world knowledge and the presence of a deontic modal in the embedded predicate or in the main one. The following and final chapter brings together the results from the previous study and this one together to answer the research questions and also discuss some of the issues which have been hinted at through these pages.

Chapter 5

Conclusions

5.1 Summary of Results

Throughout these pages, two studies on the veridicality of mood alternation and specificity have been presented and their combined results draw a fascinating but still incomplete picture.

First of all, annotations in both studies are similarly distributed: both distributions are negatively skewed and the positive labels have higher frequencies than the negative ones. Differences lay mainly in the use of the negative labels and the neutral label Uu (unknown). The first is easily explained by taking a simple look at both corpora: the pilot corpus has two mood alternation conditions (negation and possibility), whereas the main corpus comprises only the negative condition and therefore it is more likely that negative labels are used. As to the adoption of the label, Uu is explained by the same reasoning, since by having the possibility condition, which consists of adverbs like *quizás, probablemente, tal vez* (maybe, probably, perhaps), the probability of having more premise-hypothesis pairs as unknown increases greatly.

Secondly, the results of both experiments show a persistent use of the label *not a sentence*. As explained in Section 3.1, this label was included to ensure the acceptability of pairs in the possibility condition, but still, as stated in Section 3.5.1, it was used in all sorts of pairs although never with a majority agreement. Furthermore, although the possibility condition was removed from the main study, the label was kept and explicitly discouraged in the instructions. Nevertheless, as shown in Section 4.5, the label was used and in a similar proportion in comparison to the pilot study, even if with differences regarding its distribution across experimental conditions. It is difficult to point

out what exactly causes these results, but they do direct to the difficulty of the task.

Thirdly, both studies have an inter-annotator agreement score that is not quite high, especially when compared to studies like Saurí and Pustejovsky [2] or Ross and Pavlick [17]. Nevertheless, the overall agreement for the pilot study ($AC_2 = 0.484$) is still within reach of the work of de Marneffe et al. [3] ($\kappa = 0.53$), a main reference to this thesis, it is also within the range of fair agreement ($0.41 - 0.60$) [31], and thus it is not considered as problematic. Difficulties arise when we look at the negation condition in the first experiment and in the main study. On the first one, agreement was lower than the overall score ($AC_2 = 0.388$) and drops to the range of slight agreement ($0.11 - 0.40$), even if in the higher end. Although as stated in Section 3.6, it was expected that improvements in the experimental design and procedure would increase for the main study, results show the opposite: agreement decreases to $AC_2 = 0.114$, barely within the range of slight agreement. This clearly shows that the experiment design followed here is inherently harder than others [4]. A more thorough discussion on agreement is presented below in Section 5.2.

Fourthly, regarding the cumulative link mixed models (CLMMs), both show the lack of informativeness of the specificity conditions, and the most important differences between both studies lie in the variance of the random variables and on the significant effects yielded. The model for the first experiment shows considerably more variance for both annotator and pair than the main study's model. As to the significant effects, it is interesting that the first one does not have any for the negation condition but the main study's model does, even though it depends on the predictors chosen. These results suggest that there is a relevance to the veridicality of the negation condition and none to the specificity condition.

Fifthly, the analysis of the agreement patterns in the main study proved that there is indeed inherent disagreement since, as it was shown in Figure 4.4, the patterns with the highest frequencies, which are $[3, 2, 1, 1]$ and $[2, 2, 1, 1, 1]$, are separated from the others. Furthermore, distributions of these agreement patterns supported the lack of relevance of the specificity conditions since no consistent difference between them was found. As to the mood conditions, results again supported the difference between baseline and the actual mood alternation conditions, with patterns representing greater agreement for the former, and showed very small differences between the indicative and the subjunctive conditions. Also, given the patterns where indicative and subjunctive are more frequent,

it was suggested that 2 or even 3 labels represent better the pairs in these conditions.

Lastly, from the manual analysis presented in Section 4.6.5, it was learned that at least two other features have a veridicality effect on the embedded predicate: world knowledge and the presence of a deontic modal in either the main or the embedded predicate. Other possible features were suggested but with little evidence so far. Next, a more thorough discussion of the results regarding agreement is presented.

5.2 Disagreement

When gathering crowd or expert annotations it is often the case that their inter-annotator agreement score, in its different variants, is taken as a measure of how good the annotations are and thus when the expected level is not reached, different measures are taken to improve agreement. This is the path that was followed here, but given the results obtained a careful discussion on the topic is needed.

The first step is to understand what causes disagreement. Uma et al. [1] defines five sources: errors and interface problems, annotation scheme, ambiguity, item difficulty, and subjectivity. None of these can be completely rejected in our case, but given the persistence of some agreement patterns seen in Section 4.6.4 and that the interface was a simple list, not all of the disagreement can be attributed to errors or interface problems.

As to the annotation scheme, some improvements could be implemented. Firstly, stricter quality control could not just improve agreement but also yield stronger tendencies and therefore explanations. Initially, it was considered that training the annotators would prevent them from making natural annotations. But after encountering the work of Nie et al. [50], where they used carefully crafted training and testing that did not fully prevent disagreement, it is understood that training and testing could have been implemented, although it would have been a difficult implementation within the size of this corpus, the timeline given and external circumstances. In addition, the criteria for the quality control implemented were indeed defined as the annotations were encountered, rather than before running the study, and not with a lot of confidence. For example, it was clear that annotations done *too fast* had to be rejected, but the problem was defining what exactly too fast meant, and even more, what about annotations that took *too long*? Another conflicting decision was determining when labels were misused. It was

clear that if a set of annotations had been done relatively fast and had all or almost all the same label they had to be rejected. But what if they were done in a *timely manner* but still had the same label? The rejection criteria should have indeed been defined beforehand, but also some references in this matter were and are still needed. Lastly, there was a problem with the instructions. Given the feedback received from the pilot study, the instructions were reviewed before running the main experiment, but seeing the results, it is not clear whether this helped or not.

Secondly, in terms of postprocessing, there is a certain confidence in the decisions made, but there is room for improvement. On the main study pairs which were labeled as *not a sentence* were removed, but contrary to what was done in the pilot study, whole seeds were not removed. This was done to prevent losing too much data, but there is indeed no evidence to support this decision. Furthermore, a constant question when analyzing the annotations was whether more filters could be applied and how to measure their effectiveness. Thus it can be said that other filters could have improved the results.

Thirdly, given that already in the pilot study, it was proven that the effect of mood alternation depends upon the matrix verbs and, as seen in Section 4.6.3, that when not controlled it is difficult to fully analyze its effect, it is likely that a corpus balanced not just in terms of mood but also in terms of matrices could have provided more understandable results.

Lastly, after seeing the results of both studies and seeing the amount of different labels used it is clear that the assumptions must be revised. Specifically about the negation of the matrix verb. It was assumed that the negation of the matrix verb implied that the embedded predicate would be negated, even if not with complete certainty. But as seen in Section 4.6.5, this is not always the case. Thus it could have probably been better to first study the effect of negation and then of mood alternation.

Regarding the number of interpretations the relation between the different premises and their hypothesis can have, their ambiguity that is, the analyses presented in Section 4.6.4 and in Section 4.6.5 have clearly shown that there are pairs for which there are several labels that could be assigned, the decision depending on factors like what one considers to be a universal truth. This makes disagreement not an error nor a problem, but a natural occurrence. Furthermore, as already mentioned in Section 2.1, the study of Pavlick and Kwiatkowski [4] already proves the existence of an *inherent* disagreement between annotators in an NLI task, that is, they show that there are pairs for which

there is not a unique label or single truth, but rather a label split or two labels. The study of Nie et al. [50] also supports the existence of inherent disagreement.

Also within this discussion of ambiguity, there is one important topic that needs to be considered: the difference between the negation condition subset in the pilot study and the pilot subset in the main experiment. As seen in Section 4.6.2, specifically in Table 4.6, the agreement in these subsets is radically different ($AC_2 = 0.388$ and $AC_2 = 0.080$) although they consist of the exact same pairs. It is obvious that a closer examination of these annotations is needed, but at first look, they can definitely be considered as evidence that there is not always a single label that can be assigned to these pairs and more importantly, given the sociolinguistic differences between annotators (from mostly speakers from Spain to speakers from several countries) it can be considered as evidence of sociolinguistic differences regarding mood alternation which is in line with the results of Faulkner [12]. But more is needed to confirm this.

There is one last issue that should be discussed within ambiguity: the phenomena studied. The agreement scores obtained here are quite lower than previous work, but most of the previous studies mentioned here are factuality experiments and thus their corpora are not comprised of one or two phenomena, but of a set of phenomena with different levels of uncertainty, and therefore ambiguity, associated to them. But in our case, we focus on two phenomena, mood alternation, and specificity, and the first one, which has shown to have a greater effect than the second one, as already stated in Section 3.6, is a pragmatic phenomenon, and consequently a greater level uncertainty, and thus disagreement, is associated with it.

Concerning the fourth possible cause of disagreement, item difficulty, that is, how clear the interpretation of an item is, it is here a certain cause of disagreement. Firstly, as already mentioned in Section 3.6, one annotator of the pilot study mentioned that the task was difficult. Secondly, in Section 4.6.4 and in Section 4.6.5 it was explained how there are different factors to be considered when given a factuality judgment. Lastly, previous work has shown NLI annotations to be difficult [1, 4]. Thus there is ample support for the statement that the task at hand is quite difficult. To reduce it, the above-mentioned improvements could help, but it is never going to disappear entirely, maybe not even significantly.

As to the last factor for disagreement, subjectivity is also a cause for disagreement here. Although, to my knowledge, there is no previous work that supports this as cause for disagreement in NLI annotations, the analysis presented in Section 4.6.5 shows that world knowledge influences speakers' judgments, consequently making annotations dependent upon annotators' knowledge and point of view.

Now that it has been explained what caused disagreement in these studies, the question is if an inter-annotator agreement score, let it be AC_2 or any other, can reflect the nature and quality of the annotations gathered. The simple answer is no, at least not entirely. As stated in Gwet [51], inter-annotator agreement scores reflect how much the annotations change when small adjustments in the annotators are made, that is, it is a measure of data reproducibility based on the individual annotators. But given that it has been proven that these annotations are highly dependent on the speaker, measuring the reproducibility of the data based on such small variations is misguided and different evaluation scores are needed.

In this department, recent work has shown useful advances that could help better understand the annotations here gathered and increase their usability. For example, as already mentioned in Chapter 2, Pavlick and Kwiatkowski [4] computed a Gaussian Mixture Model (GMM) for the annotations of each pair. We also have the work of Nie et al. [50], which collected 100 annotations for each pair of subsets from different NLI corpora, like MNLI, computed the labels' entropy for each pair and then examined the entropy distribution for each set of annotations; and the work of Gordon et al. [52], which proposes an evaluation metric to be incorporated in machine learning algorithms that asks what proportion of the population the classifier agrees with instead of what proportion of ground truth labels the classifier agrees with. It is difficult to predict what these metrics could show about our corpora, but it is a needed and insightful exploration.

Now that a proper discussion of the results has been given, the research questions of this thesis can be answered.

5.3 Answer to Research Questions

Before trying to answer the research questions of this thesis, let us first remember what their final definition is, especially given that they were reduced from 4 to 3 after running

the pilot study:

RQ1.- In a complex sentence, how does the mood alternation of the embedded verb that occurs due to the negation of the main or matrix verb affect the factuality value of the embedded event?

RQ2.- How does an individual subject affect the factuality judgment of the event?

RQ3.- How does a subject that refers to a collective entity like an institution, affect the factuality judgment of the event?

Based on the results obtained from the two studies we can now give clearer answers than the ones initially given in Section 3.6. First, the answer for questions RQ3.- and RQ2.- is quite clear: Having an individual subject vs. having a collective barely affects the factuality of the embedded event. In other words, the specificity of the subject in terms of the number of entities to which they refer in singular is non-veridical.

As to the first question, the answer is slightly more complicated. From the results obtained in both studies we know that overall mood alternation negatively affects the factuality of the embedded predicate, but this effect is not significant. It appears though that a more detailed analysis based on matrix verbs and probably even on other factors like the presence of modal verbs will show significant effects in specific cases. In other words, the veridicality of mood alternation is bigger than that of specificity as analyzed here, but it is still rather small. Finally, let us consider what future lines of work can be developed based on what was presented in this thesis.

5.4 Future Work

Throughout these pages, two studies have been presented and their analysis focuses more on what the data looks like (AC_2, CLMMs, etc.) than on why does it look like that. Only a few possibilities, like the influence of world knowledge and the effect of modal verbs, that could explain the results have been given. Thus, a main line of future work is a thorough analysis of the data in terms of explaining how the annotations look like.

A second line of work that is probably difficult to implement but could yield very interesting results is the analysis of world knowledge. That is, instead of simply defining world knowledge as a veridicality feature in the way that it has been done here, the aim would be to understand in greater detail where it plays a greater role by, for example, studying the relation between disagreement and types of events whose factuality is questioned.

A third line of work would follow what was mentioned above about inherent disagreement and scores that better reflect what the annotations look like. For example, it could be insightful to compare both kinds of analysis to understand the differences in the information yielded by each method. In any case, gathering more annotations is required.

The fourth and last line of work proposed is the inclusion of out-of-sentence context in the corpus. The question about its inclusion was already raised while designing the main study. This seemed reasonable given the results of the pilot experiment and that, as Faulkner [12] shows, both the presence and the informativity of the out-of-sentence-context influence the acceptability of mood alternation. Furthermore, since in Section 4.6.5 it was demonstrated that the difficulties in reference resolution might have increased disagreement and uncertainty in the chosen labels, context could have solved this problem. But, given the results of Pavlick and Kwiatkowski [4] where it was shown that disagreement increased with context, it is difficult to predict how agreement would have looked like if context had been added. Furthermore, as already mentioned, Faulkner [12] shows that the informativity of the context matters, thus, adding it would have complicated an experimental design already difficult enough. Nevertheless, given the lack of studies in NLI that include context, it is definitely a line of research that needs to be developed, especially since a pragmatic approach should include context, and sentence context is not enough to fulfill this.

List of Figures

Figure 3.1 Ordered labels for the pilot study. 23
Figure 3.2 Interface for the pilot study. 27
Figure 3.3 Distribution of labels. 29
Figure 3.4 Distribution of labels by batch. 29

Figure 4.1 Interface for annotations 46
Figure 4.2 Interface with labels 46
Figure 4.3 Overall distribution of the proportion of labels used by annotators. 48
Figure 4.4 Distribution of agreement patterns. 54
Figure 4.5 Distribution of agreement patterns per mood condition. 55
Figure 4.6 Distribution of agreement patterns per mood condition. 58

List of Tables

Table 3.1 Experimental conditions and corpus generation I. 23
Table 3.2 Experimental conditions and corpus generation II. 24
Table 3.3 Label predictions I. 25
Table 3.4 Label predictions II. 26
Table 3.5 Basic counts of the annotations. 28
Table 3.6 Inter-annotator agreement scores computed for the whole dataset. . 31
Table 3.7 Inter-annotator agreement scores for experimental conditions. . . . 32
Table 3.8 Inter-annotator agreement scores for different label reductions. . . . 33
Table 3.9 Threshold coefficients. 34
Table 3.10 Assigned labels I. 35
Table 3.11 Assigned labels II. 36

Table 4.1 Experimental conditions main study. 43
Table 4.2 Some basic corpus statistics. 45
Table 4.3 Label predictions. 45
Table 4.4 Use of NaS 47
Table 4.5 Basic counts of the annotations. 49
Table 4.6 Ac2 subsets. 50
Table 4.7 AC2 matrices. 50
Table 4.8 Model Scores. 51
Table 4.9 Model Coefficients. 52
Table 4.10 Model Random Effects. 52
Table 4.11 Model Threshold Coefficients. 53

Bibliography

[1] Alexandra N Uma, Tommaso Fornaciari, Dirk Hovy, Silviu Paun, Barbara Plank, and Massimo Poesio. Learning from disagreement: A survey. *Journal of Artificial Intelligence Research*, 72:1385–1470, 2021.

[2] Roser Saurí and James Pustejovsky. Factbank: a corpus annotated with event factuality. *Language resources and evaluation*, 43(3):227–268, 2009.

[3] Marie-Catherine de Marneffe, Christopher D Manning, and Christopher Potts. Did it happen? the pragmatic complexity of veridicality assessment. *Computational linguistics*, 38(2):301–333, 2012.

[4] Ellie Pavlick and Tom Kwiatkowski. Inherent disagreements in human textual inferences. *Transactions of the Association for Computational Linguistics*, 7:677–694, 2019.

[5] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*, 2016.

[6] Christopher D Manning. Local textual inference: it's hard to circumscribe, but you know it when you see it–and nlp needs it. 2006.

[7] Roser Morante and Caroline Sporleder. Modality and negation: An introduction to the special issue. *Computational linguistics*, 38(2):223–260, 2012.

[8] Anastasia Giannakidou. (non) veridicality, evaluation, and event actualization: evidence from the subjunctive in relative clauses. In *Nonveridicality and Evaluation*, pages 17–49. Brill, 2014.

[9] Anastasia Giannakidou and Alda Mari. Mixed (non) veridicality and mood choice with emotive verbs. In *CLS 51*, 2015.

[10] Adina Williams, Nikita Nangia, and Samuel R Bowman. A broad-coverage challenge corpus for sentence understanding through inference. *arXiv preprint arXiv:1704.05426*, 2017.

[11] María Amparo Alcina Caudet. *Las expresiones referenciales. Estudio semántico del sintagma nominal*. PhD thesis, Universitat de València, 1999.

[12] Tris J Faulkner. *A Systematic Investigation of the Spanish Subjunctive: Mood Variation in Subjunctive Clauses*. Georgetown University, 2021.

[13] N. Pavlichenko, I. Stelmakh, and D. Ustalov. Crowdspeech and voxdiy: Benchmark datasets for crowdsourced audio transcription. *arXiv preprint arXiv:2107.01091*, 2021.

[14] Samuel R Bowman, Gabor Angeli, Christopher Potts, and Christopher D Manning. A large annotated corpus for learning natural language inference. *arXiv preprint arXiv:1508.05326*, 2015.

[15] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[16] Alexis Conneau, Guillaume Lample, Ruty Rinott, Adina Williams, Samuel R Bowman, Holger Schwenk, and Veselin Stoyanov. Xnli: Evaluating cross-lingual sentence representations. *arXiv preprint arXiv:1809.05053*, 2018.

[17] Alexis Ross and Ellie Pavlick. How well do nli models capture verb veridicality? In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2230–2240, 2019.

[18] Aiala Rosá, Irene Castellón, Luis Chiruzzo, Hortensia Curell, Mathías Etcheverry, Ana Fernández Montraveta, Glòria Vázquez, and Dina Wonsever. Overview of fact at iberlef 2019: Factuality analysis and classification task. In *IberLEF@ SEPLN*, 2019.

[19] Ana Fernández-Montraveta and Gloria Vázquez. The sensem corpus: An annotated corpus for spanish and catalan with information about aspectuality, modality, polarity and factuality. *Corpus Linguistics and Linguistic Theory*, 10(2):273–288, 2014.

[20] Ana María Fernández Montraveta, Hortènsia Curell i Gotor, Glòria Vázquez García, and Irene Castellón Masalles. The tagfact annotator and editor: A versatile tool. *Reproducció del document publicat a: Research in Corpus Linguistics, 2020, vol. 8, núm. 1, p. 131-146*, 2020.

[21] John Lyons. *Linguistic semantics: An introduction*. Cambridge University Press, 1995.

[22] David Sánchez-Jiménez. Una aproximación teórica a la definición del modo verbal español. 2011.

[23] RAE Real Academia Española. *Nueva gramática de la lengua española: Manual*. Espasa, 2011.

[24] José Manuel González Calvo. Sobre el modo verbal en español. *Anuario de estudios filológicos*, (18):177–204, 1995.

[25] Elisabeth Villalta. Mood and gradability: an investigation of the subjunctive mood in spanish. *Linguistics and philosophy*, 31(4):467–522, 2008.

[26] Errapel Mejías-Bikandi. Pragmatic presupposition and old information in the use of the subjunctive mood in spanish. *Hispania*, pages 941–948, 1998.

[27] Ingrid Falk and Fabienne Martin. Towards an inferential lexicon of event selecting predicates for french. *arXiv preprint arXiv:1710.01095*, 2017.

[28] Fernando García Murga. *Las presuposiciones lingüísticas*. Servicio Editorial de la Universidad del País Vasco/Euskal Herriko ..., 1998.

[29] Lucille Herrasti et al. *Características semánticas definitorias de la presuposición*. PhD thesis, El Colegio de México, 2011.

[30] Joseph L Fleiss. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378, 1971.

[31] Patrick E Shrout. Measurement reliability and agreement in psychiatry. *Statistical methods in medical research*, 7(3):301–317, 1998.

[32] Jacob Cohen. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46, 1960.

[33] Kenneth J Berry, Janis E Johnston, and Paul W Mielke Jr. Weighted kappa for multiple raters. *Perceptual and motor skills*, 107(3):837–848, 2008.

[34] Kerrie P Nelson and Don Edwards. Measures of agreement between many raters for ordinal classifications. *Statistics in medicine*, 34(23):3116–3132, 2015.

[35] Amalia Vanacore and Maria Sole Pellegrino. Robustness of κ-type coefficients for clinical agreement. *Statistics in Medicine*, 41(11):1986–2004, 2022.

[36] Matthijs J Warrens. Equivalences of weighted kappas for multiple raters. *Statistical Methodology*, 9(3):407–422, 2012.

[37] Matthijs J Warrens. Corrected zegers-ten berge coefficients are special cases of cohen's weighted kappa. *Journal of Classification*, 31(2):179–193, 2014.

[38] Anthony J Conger. Integration and generalization of kappas for multiple raters. *Psychological Bulletin*, 88(2):322, 1980.

[39] Robert L Brennan and Dale J Prediger. Coefficient kappa: Some uses, misuses, and alternatives. *Educational and psychological measurement*, 41(3):687–699, 1981.

[40] Kilem Li Gwet. Computing inter-rater reliability and its variance in the presence of high agreement. *British Journal of Mathematical and Statistical Psychology*, 61 (1):29–48, 2008.

[41] KL Gwet. Irrcac: Computing chance-correcte agreement coefficients (cac) version 1.0, 2019.

[42] Rune Haubo B Christensen. Cumulative link models for ordinal regression with the r package ordinal. *Submitted in J. Stat. Software*, 35, 2018.

[43] M. Davies. El corpus del español, 2016. URL `https://www.corpusdelespanol.org/`.

[44] Kaggle. Old newspapers, 2018. URL `https://www.kaggle.com/datasets/alvations/old-newspapers`.

[45] JS. Dario. El quijote, 2017. URL `https://gist.github.com/jsdario/6d6c69398cb0c73111e49f1218960f79`.

[46] Andreas Eisele and Yu Chen. Multiun: A multilingual corpus from united nation documents. In *LREC*, 2010.

[47] P. Gamallo, M. García, R. Martíez-Castaño C. Piñeiro, and J.C. Pichel. LinguaKit: a Big Data-based multilingual tool for linguistic analysis and information extraction. In *In Proceedings of The Second International Workshop on Advances in Natural Language Processing (ANLP 2018) co-located at SNAMS-2018*, pages 239–244, 2018. URL `http://dx.doi.org/10.1109%2FSNAMS.2018.8554689`.

[48] Stanford. Stanford lexical resources, 2012. URL `http://web.stanford.edu/group/csli_lnr/Lexical_Resources/`.

[49] Farah Benamara, Baptiste Chardon, Yannick Mathieu, Vladimir Popescu, and Nicholas Asher. How do negation and modality impact on opinions? In *Proceedings of the Workshop on Extra-Propositional Aspects of Meaning in Computational Linguistics*, pages 10–18, 2012.

[50] Yixin Nie, Xiang Zhou, and Mohit Bansal. What can we learn from collective human opinions on natural language inference data? *arXiv preprint arXiv:2010.03532*, 2020.

[51] Kilem L Gwet. *Handbook of inter-rater reliability: The definitive guide to measuring the extent of agreement among raters*. Advanced Analytics, LLC, 2014.

[52] Mitchell L Gordon, Kaitlyn Zhou, Kayur Patel, Tatsunori Hashimoto, and Michael S Bernstein. The disagreement deconvolution: Bringing machine learning performance metrics in line with reality. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–14, 2021.

Appendix A

Additional Scores

A.1 Pilot Study

A.1.1 Inter-Annotator Agreement Scores

Agreement Scores	
Name	Value
Fleiss' κ	0.160
Fleiss' κ with linear weights	0.177
Conger's κ with linear weights	0.176
Gwet's AC_2 with linear weights	0.484

Agreement Scores for Experimental Conditions			
Subset	κ_w^f	κ_w^c	AC_2
ALL	0.177	0.179	0.484
Negation	0.094	0.097	0.388
Possibility	0.278	0.278	0.592
Individual	0.161	0.162	0.500
Collective	0.190	0.192	0.470

Agreement Scores for Different Combinations of 3 Raters			
Combination	κ_w^f	κ_w^c	AC_2
ALL	$0,177$	$0,179$	$0,484$
R1, R2 and R3	0.170	0.172	0.501
R1, R2 and R4	0.237	0.239	0.550
R1, R2 and R5	0.219	0.223	0.493
R1, R3 and R4	0.181	0.182	0.527
R1, R3 and R5	0.177	0.180	0.478
R1, R4 and R5	0.242	0.246	0.527
R2, R3 and R4	0.133	0.134	0.469
R2, R3 and R5	0.099	0.102	0.396
R2, R4 and R5	0.162	0.165	0.447
R3, R4 and R5	0.147	0.150	0.450

Agreement Scores for Different Label Reductions			
Reduction	κ_w^f	κ_w^c	AC_2
ALL	$0,177$	$0,179$	$0,484$
No NaS	0.190	0.191	0.469
PS with PR	0.186	0.187	0.572
PS with PR, no NaS	0.204	0.205	0.537
PR with CT, PS with Uu	0.153	0.155	0.635
PR with CT, PS with Uu, no NaS	0.162	0.164	0.518
PS with Uu	0.204	0.205	0.590
PS with Uu, no NaS	0.224	0.225	0.550

A.1.2 Cumulative Link Mixed Model

Model Scorers							
link	Threshold	Nobs	logLik	AIC	niter	max.grad	cond.H
logit	flexible	1530	-2456.68	4961.36	4502(13559)	5.68×10^{-3}	8.5×10^2

Model Coefficients				
Coefficient	Estimate	Std. Error	z value	$\Pr(>\|z\|)$
moodconadverb	2.151	0.282	7.632	2.31×10^{-14}
moodcatindicative	0.427	0.311	1.371	0.170
moodcatsubjunctive	−0.063	0.305	−0.206	0.836
specificityconcollective	−0.199	0.250	−0.797	0.425
specificitycatmixed	0.280	0.329	0.851	0.395
specificitycatproper	0.024	0.328	0.072	0.9426
moodconadverb:moodcatindicative	−3.028	0.334	−9.072	$< 2 \times 10^{-16}$
moodconadverb:moodcatsubjunctive	−2.868	0.330	−8.676	$< 2 \times 10^{-16}$
specificityconcollective:specificitycatmixed	−0.291	0.311	−0.933	0.351
specificityconcollective:specificitycatproper	0.107	0.312	0.344	0.731
moodconadverb:specificityconcollective	0.494	0.256	1.928	0.054
moodcatindicative:specificitycatmixed	−0.209	0.390	−0.536	0.592
moodcatsubjunctive:specificitycatmixed	−0.366	0.387	−0.947	0.344
moodcatindicative:specificitycatproper	−0.346	0.391	−0.884	0.377
moodcatsubjunctive:specificitycatproper	−0.131	0.387	−0.338	0.735

Model Random Effects			
Groups	Name	Variance	Std.Dev.
PAIR	Intercept	0.512	0.715
RATER	Intercept	0.505	0.710

Model Threshold Coefficients			
Threshold	Estimate	Std. Error	z value
NaS—CT-	−4.254	0.326	−13.062
CT-—PR-	−3.272	0.298	−10.962
PR-—PS-	−2.756	0.291	−9.480
PS-—Uu	−2.371	0.286	−8.277
Uu—PS+	−0.904	0.278	−3.249
PS+—PR+	0.006	0.277	0.023
PR+—CT+	1.233	0.279	4.415

A.2 Main Study: CLMMs with Most Frequent Matrices

A.2.1 Predictors: Mood

Model Scorers							
link	Threshold	Nobs	logLik	AIC	niter	max.grad	cond.H
logit	flexible	1701	−3125.09	6270.18	1024(2105)	5.04×10^{-3}	1.4×10^{2}

Model Coefficients				
Coefficient	Estimate	Std. Error	z value	Pr($>\|z\|$)
MOOD_CONDITIONindicative	−0.5769	0.1156	−4.990	6.05×10^{07}
MOOD_CONDITIONsubjunctive	−0.6338	0.1161	−5.462	4.72×10^{08}

Model Random Effects			
Groups	Name	Variance	Std.Dev.
PAIR	Intercept	0.080	0.283
RATER	Intercept	0.014	0.120

Model Threshold Coefficients			
Threshold	Estimate	Std. Error	z value
CT-—PR-	−3.075	0.126	−24.447
PR-—PS-	−2.100	0.102	−19.769
PS-—Uu	−1.383	0.094	−14.741
Uu—PS+	−0.888	0.090	−9.905
PS+—PR+	−0.200	0.087	−2.308
PR+—CT+	0.560	0.088	6.376

A.2.2 Predictors: Mood and Matrix

Model Scorers							
link	Threshold	Nobs	logLik	AIC	niter	max.grad	cond.H
logit	flexible	1701	−3112.90	6269.81	3528(7150)	5.04×3.19^{-3}	2.6×10^3

Model Coefficients			
Coefficient	Estimate	Std. Error	Pr(> |z|)
MOOD_CONDITIONindicative	−0.699	0.297	0.019
MOOD_CONDITIONsubjunctive	−1.309	0.305	1.76×10^{05}
MATRIXconsiderar	−0.404	0.292	0.167
MATRIXcreer	−0.584	0.280	0.037
MATRIXdecir	−0.139	0.312	0.656
MATRIXsaber	−0.053	0.252	0.834
MOOD_CONDITIONindicative:MATRIXconsiderar	0.252	0.405	0.534
MOOD_CONDITIONsubjunctive:MATRIXconsiderar	0.807	0.401	0.044
MOOD_CONDITIONindicative:MATRIXcreer	0.154	0.390	0.693
MOOD_CONDITIONsubjunctive:MATRIXcreer	0.836	0.393	0.033
MOOD_CONDITIONindicative:MATRIXdecir	−0.125	0.422	0.768
MOOD_CONDITIONsubjunctive:MATRIXdecir	0.677	0.441	0.125
MOOD_CONDITIONindicative:MATRIXsaber	0.156	0.345	0.652
MOOD_CONDITIONsubjunctive:MATRIXsaber	0.767	0.352	0.029

Model Random Effects			
Groups	Name	Variance	Std.Dev.
PAIR	Intercept	0.028	0.167
RATER	Intercept	0.014	0.120

Model Threshold Coefficients			
Threshold	Estimate	Std. Error	z value
CT-—PR-	−3.289	0.239	−13.781
PR-—PS-	−2.225	0.227	−9.804
PS-—Uu	−1.600	0.223	−7.157
Uu—PS+	−1.105	0.222	−4.987
PS+—PR+	−0.417	0.220	−1.899
PR+—CT+	0.343	0.219	1.564